7

Connectionist explanation: taking positions in the mind–brain dilemma

Paul F.M.J. Verschure
Institute for Neuroinformatics
ETH/University of Zurich

7.1 Introduction

The computer metaphor of cognitivism that has had such a strong influence on cognitive science over recent decades seems to be confronted (again) by a competitor: the brain metaphor put forward by connectionism (e.g. [McClelland and Rumelhart 1986] and [Sejnowski *et al.* 1988]). Connectionism assumes that mental phenomena can be explained in terms of the parallel activation and interaction of a large number of units (model neurons). These units are linked by connections (artificial synapses) which modulate the transmitted activity. Knowledge is represented in these connections between the units and learning takes place by adjusting their strength. An important, and much emphasized, aspect of connectionist models is their emergent behaviour. The massive parallel interaction of a large number of simple units can lead to qualitatively different and more interesting forms of behaviour. Successes of this connectionist approach range from models of human memory (e.g. [McClelland and Rumelhart 1986]) to practical applications that can control the navigation of cars [Pomerleau 1989]. An interesting property of these models is their robustness against loss of interconnections. This is interpreted by some as an indication of their biological plausibility (e.g. [Hinton *et al.* 1991]).

An important topic in the discussion about the relevance of connectionism for cognitive science has become the issue of levels [Estes 1988]: must connectionist models be interpreted at the level of physical instantiation

or at the level of symbol manipulation? Critics of the connectionist movement (e.g. [Broadbent 1985] and [Fodor and Pylyshyn 1988]) argue that it cannot be considered an alternative to the classical cognitivist paradigm. They characterize connectionism as an attempt to define a brain-like implementation of symbol manipulating models. Supporters of connectionism, however, emphasize that it is an alternative paradigm in the study of cognition that will largely contribute to our understanding of the mind–brain duality. [Smolensky 1988], for instance, claims that connectionism is the most significant development in the philosophy of mind over past millennia.

In contrast to the connectionist movement of the 1950s an important objective of present-day neo-connectionism is now the definition of a *subsymbolic* bridge between the formal mind, as studied in cognitive science, and the brain, as studied in neuroscience. This objective implies that subsymbolic connectionism has to address the question of how the symbolic description of psychological processes, in terms of rules and representations, provided by traditional cognitivism can be related to a non-symbolic one in terms of brain mechanisms. [Haugeland 1978] has called this *the problem of complete reduction*. A substantial contribution to this connectionist ambition has been made by [Smolensky 1987, Smolensky 1988], who has tried to explicate the subsymbolic connectionist paradigm.

I will evaluate the solution of the problem of complete reduction proposed by subsymbolic connectionism by analysing NETtalk, which is often taken as the paradigmatic example of this approach, and of a closely related network model that can classify sonar targets. To further generalize the results of this analysis a generic connectionist model, the autoassociator, will be examined.

The results presented suggest that subsymbolic connectionism, as defined by the models analysed, is still completely dependent on the symbolic level of description and does not live up to its promise of bridging the gap between mind and brain. Therefore, this approach is facing the same problems that confront a cognitivistic approach.

This result can be interpreted as a confirmation of the hegemony of a symbolic approach. I, however, prefer to propose an alternative interpretation, which is directed at defining the circumstances under which the initial ambition of connectionism, to find relations between symbolic characterizations of behaviour and biological ones, can be realized. The background of this proposal is not a personal deficit on the part of the author to disagree with tradition. It is based on the conclusion that the cognitivistic school of thought is confronted with a series of fundamental problems. The proposed alternative focuses on two questions:

1. What is the nature of connectionist models and what are the criteria that

should be applied to them to render them plausible models of cognition and behaviour?

2. What should be the relationship between the dynamical view on cognition inherent in connectionist models and a symbolic one?

These questions bring us back to classical epistemological issues: what is knowledge and where does it come from? The two questions raised will be used to build up an argument to address the more fundamental ones. The answer to the first question can be seen as the definition of a research strategy for connectionism while the answer to the second one will provide a conceptual framework. The background of this discussion will be the contrast between cognitivism and traditional connectionism: the mind–brain dilemma.

7.2 The mind–brain dilemma

The paradigm that has dominated cognitive science over the last few decades can be called 'symbolic cognitive psychology' [Newell 1990]. This approach bases its explanations of behaviour on an assumed knowledge level. The laws which explain behaviour relate knowledge to goals according to the principle of rationality: roughly, a system will use its knowledge to reach its goals. 'If the system wants to attain goal G and knows that to do act A will lead to attaining G, then it will do A. This law is a simple form of rationality – that an agent will operate in its own best interest according to what it knows' (op. cit. p. 49).

In their original proposal [Newell and Simon 1963] intended to define a paradigm for explaining and studying intelligence that would relax the restriction put forward by behaviourism that behaviour should be explained in terms of observable events without referring to mediating processes. Moreover, they tried to avoid the use of subjective notions in the explanation of behaviour as was done in the school of phenomenology whose main method was introspection.

The empirical hypothesis put forward by this approach is that general intelligence, as defined at the knowledge level, can only be displayed by systems that can manipulate symbols: *physical symbol systems* [Newell and Simon 1976, Newell 1990]. General intelligence here is taken to mean that the system is supposed to operate in domains where 'within some broad limits anything can become a task' [Newell 1990]. This hypothesis specifies the research programme of AI and can be seen as the paradigmatic example of cognitivism. The hypothesis states that a physical symbol system (PSS) constitutes the necessary and sufficient conditions for general intelligence. A PSS is embedded in an environment that consists of discrete states: objects and their relations. The system itself possesses a memory, a set of operators, control, input and output. The output of the system is a

function of the input. This response function is defined by the memory, the operators and the control. The memory of the system consists of a number of symbolic expressions. Symbols are taken to be patterns in a physical medium that stand for properties of the physical world. Symbols can again be combined to form expressions. Expressions are stable and will persist in the system until they are explicitly changed by some operator. Symbols designate states inside or outside the system by allowing the system to access additional information about these states. They acquire their meaning by being interrelated with other symbols. To access additional information implies that the system has to perform a search. What a symbol designates is not given in advance. The content of an expression, however, is defined by the designation of the constituting symbols. The operators take symbols or expressions as input and produce one or more symbols as output. The control governs the behaviour of the system by interpreting symbolic expressions: the system will carry out the processes designated by the interpreted expression. Symbols can represent objects and their relations since they can designate the relevant information. The connection from symbols to states of the outside world is established by transduction rules that map sensory states, which relate to states of the world, onto internal symbolic representations of these states.

The basic rules that specify what actions the agent can generate are specified by a finite set of logical axioms implemented by the symbols and operators of a PSS. The knowledge of the agent becomes the logical closure of these axioms. Although the knowledge-level description of the behaviour of an agent can never be complete it allows an observer to make predictions about its behaviour if the knowledge and goals are known. Newell, and others, assume that the best we can do in the explanations we give of intelligence, and for that matter of our own behaviour, is to try to approximate the knowledge level.

Knowledge-level explanations of intelligence and the notion of physical symbol systems are seen as the result of decades of work in AI and computer science: 'it is the structure of the digital computer itself (and the theoretical analysis of it) that reveals the nature of symbolic systems' [Newell 1990]. It should not be taken as a computer metaphor but as a scientific theory of the qualitative structure of the mind which is supposed to have the same status as, for instance, the cell doctrine in biology.

The analogy with the way in which we interact with computers is obvious. To understand or define the operation of a computer we mostly limit ourselves to the logic of the algorithm that is executed and not to the actual activity taking place in the hardware components. It is assumed that in order to understand cognition knowledge of the algorithms that the brain is supposed to execute is all that matters. This implies that we have to drop the ideal that there is a unity of science in which all sciences are in principle reducible to physics (e.g. [Putnam 1960, Fodor 1974, Haugeland 1978]). It

is assumed that there is one 'special science' of cognition which develops a formal/logical description of the rules and representations constituting it. The level of implementation is considered of minor importance.

[Winograd and Flores 1986], p.15, describe the basic steps followed in this paradigm to explain cognition as:

> 1. Characterize the situation in terms of identifiable objects with well defined properties.
> 2. Find general rules that apply to situations in terms of those objects and properties.
> 3. Apply the rules to the situation of concern, drawing conclusions about what should be done.

An influential example of this tradition is the General Problem Solver proposed by [Newell and Simon 1963] which was not only capable of solving difficult (logical) problems but also seemed to solve these problems in a human-like way. More recently [Newell 1990] has proposed the more advanced SOAR architecture as a possible unifying theory of cognition.

This view of cognitive science as a special science has led to an extreme stance on the role that (neuro)biology plays in explaining behaviour. Since algorithms that are expressed in computer programs can be instantiated in an arbitrary number of different implementations, it follows that we do not have to bother with the hardware in which mental software happens to be executed: the brain. Basically, this way of dealing with the problem of complete reduction comes down to asserting that intelligence can be achieved without having a brain. Cognitivism limits itself to the identification of the logic of the rational mind. The problem of complete reduction is no longer relevant since the physical instantiation of cognition in the brain does not constitute a level of analysis that poses constraints on the symbolic and knowledge level of analysis that matters.

The physical symbol systems hypothesis is often seen as the only plausible model of general intelligence and having no serious competitors. One of its virtues, regularly emphasized by its proponents (e.g. [Pylyshyn 1989]), is that it gives us a theory to understand complex cognitive functions like language and reasoning. In fact Chomsky's theory of transformational grammars was an important impetus to the development of cognitivism (see [Gardner 1987] for a historical overview). We cannot forget, however, that this approach also has its problems, such as the implication of nativism, the related problem of symbol grounding, the frame problem, the frame of reference problem, and the problem of situatedness. As a background for the present discussion these problems will be briefly described.

The first problem of this position is its implied *nativism*. By assuming that cognition arises out of the manipulation of rules and representations one must explain where these primitives come from. In the case of a computer application we can always refer to a human programmer. In the case of cognitivism, where a reference to a programmer would be rather

awkward, one is forced to blame evolution: nativism. It is important to note that it is not a matter of choice whether one wants to marry a cognitivistic orientation with nativism. The choice for cognitivism forces one to submit to this implication.

[Newell and Simon 1976] state that if one thinks that there is nothing problematic or mysterious about physical symbol systems and the way in which they explain intelligence one is a child of today (today in this context means the 1950s, 1960s and 1970s). Moreover, they contrast their solution with the classical one of Plato, where problem solving was seen as recalling experiences from a previous occurrence. Newell and Simon claim that this 'preposterous' solution can now be replaced by a much simpler one. For a student of today (and now I mean today), however, their solution does not seem that different from Plato's proposal and it is certainly not unproblematic.

If we take a closer look at the models developed in this paradigm we see that they derive all their problem-solving behaviour from their predefined body of knowledge. This implies that the full burden of explaining intelligence is placed on the question as to where this predefined body of knowledge comes from. For Plato it was a previous life and for these systems it is the creator of the program. For the explanatory value, it makes little difference. Because of the assumption behind physical symbol systems that they possess a complete body of knowledge from the start on, their proponents are forced into a nativist position where all of these properties of the system, down to the last symbol, are genetically predefined.

This implication, however, leads to two types of problems. The first is of a practical nature and the second of a principled one. By residing to nativism one assumes that the genetic code is able to store the rules and representations necessary for the construction of a cognitive engine. Moreover, it must be able to translate this into a very precisely orchestrated process of morphogenesis, in which billions of cells are involved. Nativism supposes that the genetic code is able to generate very precise wiring schemes between, for instance, the sensory systems and memory structures to assure the reliable transduction of sensory states into internal representations. The practical problem is that the genome does not have the necessary amount of genes to be able to accomplish this job [Changeux 1985, Edelman 1987]. Of the some 50 000 genes involved in realizing the whole vertebrate phenotype only around 30 000 correspond to the brain. That they are expressed in the brain, however, does not mean that they are all necessary to develop one [Miklos 1993]. This is one of the reasons to assume that the brain is not constructed according to nativist principles, leading to very precise and predefined point-to-point wiring, but on selectionist ones exploiting the basic principles of the generation of diversity and selection by means of differential amplification [Changeux 1985, Edelman 1987].

If one, in the face of this first problem, would still like to insist on the

implied nativism of cognitivism, one must still explain how, during evolution, all this 'knowledge' could have accumulated in our genes (e.g. Piaget in [Piatelli-Palmarini 1980]). Some might argue that cognitivism has never denied learning, and also has proposed certain interesting learning mechanisms, like Newell's chunking. A closer analysis of this issue, however, reveals that this form of learning has nothing to do with acquiring 'knowledge' from the interaction with the world. Learning in this case has been predefined. It exclusively focuses on how the search in a predefined body of knowledge, the closure of the logical axioms underlying the rules and representations of a PSS, can be reduced to a subset of all possible courses of solution by means of storing successful ones; that is on chunking. The role of learning has become one of optimizing search in a predefined model of the world instead of acquiring knowledge from interacting with the real world. (See [Verschure and Pfeifer 1993] for a further analysis.) Although cognitivists might not deny that organisms learn and adapt to their environment, there is no place for this phenomenon in the explanation of behaviour they offer.

The *symbol-grounding problem* [Searle 1980, Harnad 1990] deals with the assignment of meaning to symbolic representations.[1] In systems developed in this tradition the meaning of symbols is derived from the way they are connected to other symbols and how they are processed. An important assumption in the explanation offered by cognitivists is the transduction function that transforms sensory states into symbolic states. Transduction establishes the connection between the real world and the internal world model. Newell admits that in the cognitivistic approach the meaning of symbolic representations is not explained but is assumed [Newell 1981, p. 18]. Moreover, it is taken for granted that sensors are somehow capable of reliably relating, or transducing, events in the world to their related internal symbolic representations.

The *frame problem* [McCarthy and Hayes 1969, Pylyshyn 1987] indicates that it is impossible to maintain a symbolic world model of a complicated environment while acting in it in real time. Since the time needed to update this world model will increase exponentially when it becomes more extended, the system, at a certain point, will become completely absorbed in maintaining it. This in turn will prevent it from acting. Originally this problem was defined in relation to the difficulty of drawing the right inference, given a logical representation. Recently it has received a more general interpretation (e.g. Janlert in [Pylyshyn 1987]). It is important to realize that the frame problem cannot be solved by relying on increasing the

[1] The symbol-grounding problem is also used as a denominator for the class of problems relating to language and connectionist modelling. In this discussion I refer to a more restricted but also more principled problem as defined by Searle and Harnad: how can a symbolic system have a sense of meaning?

computing power. It relates to the principled impossibility of maintaining a consistent symbolic model of a dynamically changing world.

Essentially the *frame-of-reference problem* [Clancey 1989, Clancey 1992] conceptualizes the relation between the observer, the designer (or the modeller), the artifact (e.g. the expert system, the robot), and the environment. Let us take the example of building a classical expert system. The 'knowledge engineer' is at the same time observer and designer. He or she observes and describes in symbolic terms the problem-solving behaviour of a human expert. This implies the definition of a domain ontology, that is a categorization of the real-world domain by the knowledge engineer (of course, based on his or her interaction with the expert). This ontology is taken as the basis for the system development. This has a number of consequences. First, the categories and the symbols used for them are the ones of the knowledge engineer: they are grounded in his or her experience, not in the experience of the system. Second, this domain ontology is static.

These consequences in turn lead to some problems. Given that the world is continuously changing, a static ontology will always at some point become inappropriate, that is the system will not be adaptive. If a situation is encountered which cannot be appropriately captured by the ontology of a system, its behaviour will be inadequate. Even if the system can learn, the primitives and therefore the classes of objects and events will remain the same. Therefore, if a system is to interact successfully with its environment it must be able to form its own classification. This classification must be capable of continuously adapting to change. It must be embedded in the properties of the system–environment interaction.

If we take the human, say the user of a system, out of the previously described design loop the symbol-grounding problem becomes relevant: since the symbols are not grounded in the system's experience it will not be able to make the connection to the outside world. Especially for the field of autonomous robots, where the system has to interact with the environment without the intervention of a human, this poses a problem. Since the symbols are observer- or designer-based it is not clear, however, why the system should contain symbols in the first place. Indeed, the symbol-grounding problem can be taken as an artifact of the symbolic approach.

The *situatedness* of intelligent systems indicates that they have to deal with a constantly changing, partially unknowable, and unpredictable world [Simon 1969, Winograd and Flores 1986, Agre and Chapman 1987, Suchman 1987]. These systems have to act in real time since the environment is constantly changing, largely – but not only – because of what other agents do. The traditional symbolic approach to designing agents is to equip them with models of their environment. These models form the basis for planning processes which in turn are used for deciding on a particular action. But plan-based agents very quickly run into combinatorial problems (e.g. [Chapman 1987]) because in an unpredictable world many alterna-

tives must be considered. Since the environment is only partially knowable a complete model cannot be built in the first place. However, even if only partial models are developed, keeping the models up to date requires a lot of computational resources. This pertains to the frame problem that was discussed earlier. Inspection of the problem of taking action in the real world shows that it is neither necessary nor desirable to develop 'complete' and very detailed plans and models (e.g. [Suchman 1987], [Winograd and Flores 1986] and [Verschure and Pfeifer 1993]).

The issue of situatedness indicates that there is no simple way around the symbol-grounding problem. In fact, cognitivism is confronted here with one of the problems that Chomsky raised against behaviourism. In his famous review of Skinner's theory of language, [Chomsky 1959] shows that, for instance, the reference to stimulus control made by behaviourists is quite meaningless outside the restricted set-up of laboratory experiments. It completely rests on an *a posteriori* interpretation of the response by an observer, in which the relevant stimulus properties are defined. If a person would utter 'red' after seeing a red chair, the theory of stimulus control would say that this behaviour is under the control of the stimulus property 'redness'. If, however, the person would have said 'chair' the relevant stimulus property would have been 'chairness'. Any interpretation becomes possible in terms of stimulus control. The meaning of the notion 'stimulus' has lost all the objectivity it is supposed to have in the behaviourist tradition. It is no longer part of the outside physical world but is a construct ascribed to the system by the observer. Therefore, predictions about behaviour cannot be made any more. Chomsky sees this as a retreat to mentalistic psychology which relies on a general mystification. Cognitivism, however, does not seem to fare much better than behaviourism in this respect. In the cognitivist tradition the rules and representations used in explaining a response function are *a priori* ascribed to the system by the designer; they are only connected to the outside world in the frame of reference of the observer or designer. Therefore, Chomsky's argument was not only a serious blow to behaviourism, but also specified a problem that cognitivism has not even started to solve.

I will not embark on a full evaluation of the cognitivist tradition, but simply note that one pays a price for the conceptual conveniences offered by the knowledge level and the hypotheses of physical symbol systems put forward by this tradition. It still has to solve some serious problems. Some of these problems, however, seem to be a direct implication of the assumptions underlying this paradigm in the first place. These considerations form a starting point for the work that will be presented later. The conceptual issues discussed, however, indicate that there is no reason to believe that the cognitivistic paradigm is the only game in town. It faces some deep problems which render it not totally convincing. This can be seen as an invitation to explore alternatives.

7.3 Connectionism

Another influential approach in cognitive science is connectionism (e.g. [Rosenblatt 1958]). Instead of relying on an analogy with digital computers, its theorizing about cognition is closely tied to knowledge about the brain (e.g. [Hebb 1949]). Rosenblatt assumes that by interacting in its environment an organism, which does not possess prior knowledge of this environment, develops preferences for specific responses to certain stimuli. The developing associations between stimuli and responses are related to the ramification of distinct connection patterns in its nervous system. The classical example of this approach is the perceptron proposed by [Rosenblatt 1958, Rosenblatt 1962], who tried to develop a formal theoretical basis for the study of biological intelligence. Instead of relying on symbolic logic Rosenblatt founded his explanations in probability theory. In this approach the problem of complete reduction is dissolved by rejecting a description at the level of logical symbol manipulation and strictly relating cognition to its substrate: the brain.

In the days of Rosenblatt it was seen as an advantage of this approach that it allowed psychology to stay in touch with (neuro)biology. The phenomena that could be modelled in connectionist systems, like pattern recognition, however, seemed far removed from the phenomena that should be explained by cognitive science. The relation of this approach to associationism, which in that period was seen as a hopeless school of thought (see for instance the already discussed criticism of Chomsky) was seen as indicating that these techniques would not suffice to explain phenomena like language [Bechtel 1989]. Moreover, [Minsky and Papert 1969] demonstrated that the perceptron was not able to perform certain computations. These results were interpreted as undercutting the theoretical thrust of connectionism and attention died away (or failed to increase), leaving cognitive science in the firm grip of cognitivism for quite some time.

If we consider the two main schools of thought in cognitive science it appears that this field is confronted with a mind–brain dilemma. It seems that a complete understanding of the mind and the brain can only be found if either the (formal) mind is left out, as proposed by connectionism, or the brain is left out, as is suggested by cognitivism. The recently proposed approach of subsymbolic connectionism, however, wants to solve the mind–brain dilemma by trying to reconcile the two seemingly orthogonal approaches.

7.4 Cognition and subsymbols

Although the models proposed by neo-connectionism are strongly related to Rosenblatt's perceptron, their theoretical context is completely different. The recent attempt by [Smolensky 1988] to define an alternative theoretical

framework for connectionism is not as radical as Rosenblatt's proposal. Smolensky emphasizes the importance of a subsymbolic approach towards understanding cognition, which mediates between the formal mind and the dynamical brain.

The subsymbolic paradigm, which was initially proposed by [Hofstadter 1985], is based on developments in the present mainstream of connectionist research (e.g. [McClelland and Rumelhart 1986]). In this proposal the rules and representations of cognitivism are seen as emergent properties of the interaction of a large number of subsymbolic units. Symbols are encoded by the 'complex patterns of activity over many units. Each unit participates in many such patterns. The interactions between individual units are simple, but these units do not have conceptual semantics: they are subconceptual' [Smolensky 1988, p. 6].

The subsymbolic description of cognition is supposed to be, in principle, reducible to brain processes. Or, as Smolensky puts it, 'if we succeed in building symbols and symbol manipulation out of connectoplasm then we will have an explanation of *where symbols and symbol manipulation come from* [...] With any luck we will even have an explanation how the brain builds symbolic computation' ([Smolensky 1987, p. 141], emphasis in original). The limited knowledge we have of the central nervous system is seen as the only obstacle to be overcome towards finding this subsymbolic explanation of cognition.

The explanation of cognition offered by subsymbolic connectionism is characterized by a dimension shift from a symbolic description at the level of complex patterns of activity to a non-symbolic one at the level of subsymbolic units. Smolensky, however, emphasizes that 'for the time being, subsymbolic models of higher processes are much more directly related to conceptual level accounts of these processes than to any neural account' [Smolensky 1988, p. 8]. Like others, Smolensky compares this scheme of explanation, subsymbolic reduction, with that of physical science. Just as classical mechanics gives a useful and accurate higher-level description of the interaction of macroscopic bodies, cognitivism gives a useful and accurate description of macroscopic cognitive processes. The complete descriptions of these processes, however, have to be found at a quantum-mechanical or subsymbolic level, respectively.

Other theorists have advocated alternatives to Smolensky's account of the relation between the symbolic and the subsymbolic level of description. [Churchland 1989] and [Ramsey *et al.* 1991], for example, argue that the subsymbolic explanation of a task performed by a connectionist model should not be located at the level of the units (where Smolensky places it) but at the level of the formal laws that govern the behaviour of the model (the equations determining the evolution of weights and activation patterns in time) and/or the weights connecting the units. This alternative, however, endorses the same claim: connectionist models are capable

of representing and manipulating information in a qualitatively different way than symbolic models.

Subsymbolic connectionism proposes an alternative position to both those of cognitivism and traditional connectionism on the mind–brain dilemma by postulating that a third level of description, in between the formal mind and the dynamical brain, is essential for understanding cognition. This claim of a subsymbolic level distinguishes subsymbolic connectionism from cognitivism. In assuming that cognitivism constitutes a useful, though incomplete, level of description, subsymbolic connectionism distinguishes itself from traditional connectionism (as envisioned by, for instance, Rosenblatt). Where the latter tried to define an anti-cognitivistic alternative, the subsymbolic connectionism tries to relate the two approaches through the subsymbolic level of description.

If subsymbolic connectionism proves to be correct, it could indeed be interpreted as important progress in the philosophy of mind. It would show a way out of the mind–brain dilemma by unifying the rational mind with the dynamical brain. This justifies a closer evaluation of the power of subsymbolic computing. By examining the way NETtalk, and a closely related model for the classification of sonar return signals, represent their task domains, I will try to assess the credibility of the subsymbolic promise. With another example, a connectionist model of Donald Duck, I will evaluate the claims made by Churchland and Ramsey *et al.*

7.5 The power of subsymbolic computing

A standard example of a connectionist model that displays interesting emergent behaviour is NETtalk (the famous 'parallel network that learns to read aloud') developed by [Sejnowski and Rosenberg 1986]. Proponents of subsymbolic connectionism assume that the hidden units (the units between the input and output layers) in connectionist models like NETtalk exhibit subsymbolic representations and thus illustrate the power of subsymbolic computing. Although the designers of NETtalk acknowledge the differences between the architecture of NETtalk and the brain, they assume that NETtalk can teach us how information (in this case letter-to-phoneme mappings) could be represented in 'large populations of neurons' [Sejnowski and Rosenberg 1986, p. 670]. [Churchland and Sejnowski 1989, p. 244] indicate that it 'yields clues to how the nervous system can embody models of various domains of the world'.

With NETtalk Sejnowski and Rosenberg have quite successfully modelled the conversion of English text to speech. NETtalk is proposed as a connectionist alternative to DECtalk, a commercial product that was designed for this task based on symbolic techniques. Two major functions must be carried out in order to make this conversion. First, the text must be mapped into an abstract linguistic description consisting of phonemes, stress and

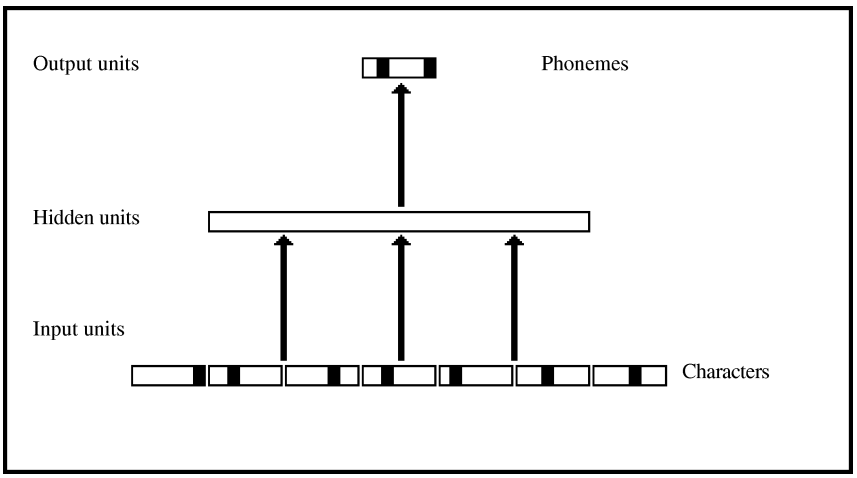

Figure 7.1 *The architecture of NETtalk: see text for explanation*

syntactic information. Second, this linguistic description must be mapped into synthetic speech by translating the phoneme string, along with lexical stress, syntactic and semantic information, into an acoustic wave form. This linguistic description is defined in DECtalk as a set of articulatory features that specify the parameter values for a formant speech synthesizer. NETtalk was designed to perform the mapping from the text to the articulatory features directly using the same coding as applied in DECtalk.

The architecture of NETtalk consists of three layers: input, hidden and output (Fig. 7.1). The input layer of NETtalk contains 7 identical groups of 29 units each. Each unit in a group codes a letter of the alphabet, a word boundary, or punctuation. The hidden layer has no preassigned interpretation but is necessary to accomplish the mapping between the input and the output layers. Each unit of the output layer represents one of 23 articulatory features or one of 3 features representing stress and syllable boundaries. Hence, on the input and output layers, all features are represented locally by single units.

The network learns to associate the letter coded by the fourth group of the input layer with a specific set of pronunciation features represented by the output layer. The other 6 groups of the input layer provide a context. To learn the pronunciation of a letter, a specific pattern of activation of the input layer, representing the letter and the context, must be associated with a pattern of activation of the output layer (depicted by the black squares in the figure). Learning proceeds in a supervised way: the weights connecting the units of the input, hidden and output layers will be adjusted to reduce the difference between the pattern of activation actually generated in the

output layer, due to the received input from the hidden units, and what the pattern actually should be, as specified by the training set.

NETtalk was able to learn the associations between letters and phonemes and could correctly pronounce 95% of the presented words after training with 50 000 words. It could correctly generalize to new cases for 78% of the presented test words.

[Rosenberg and Sejnowski 1987] (see also [Rosenberg 1986]) tried to determine the features coded for by the hidden units of a trained network by clustering input patterns that led to similar activation patterns of these elements. The cluster analysis of NETtalk showed that the activity patterns of the hidden units could be categorized into two main groups: vowels and consonants. These results were considered to be an important proof of the power of subsymbolic computing: they demonstrated the emergence of a 'symbolic' separation of the letter-to-phoneme mapping into vowels and consonants.

According to [Sejnowski and Rosenberg 1986] NETtalk started out with no 'considerable innate knowledge in the form of input and output representations that were chosen by the experimenters'. Several of the important results of this work, as summarized by [Churchland and Sejnowski 1989], are:

1. 'The representational organization [of the network] is not programmed or coded into the network; it is found by the network. In a sense it "programs" itself' [p. 239].

2. 'The representation is a property of the collection of hidden units, and does not resemble sentence-logic[2] organization' [p. 239].

3. '[Networks that start out with] different initial conditions [...] had similar functional clusterings [the vowel–consonant distinction]' [p. 239].

This brings them to conclude that network models like NETtalk show 'how knowledge of brain architecture can contribute to the devising of likely and powerful algorithms that can be efficiently implemented in the architecture of the nervous system and may alter even how we construe the computational problems' (p. 246).

NETtalk is surrounded with an impressive set of interpretations and claims. Again, in case this all proves to be correct, we are witnessing an important moment in cognitive science. Given this importance a closer analysis of this model is necessary. Since the main claim is the emergence of a vowel–consonant distinction, this regularity should not be present in the character–phoneme relationships expressed in the input and output patterns presented to NETtalk. In their article Sejnowski and Rosenberg

[2] For Churchland and Sejnowski, the expression 'sentence-logic' refers to a symbolic approach.

Table 7.1 *Vowels in NETtalk*

Phoneme (example)	Articulatory features				
a (f**a**ther)	Central 2			Low	Tensed
c (b**ough**t)	Velar			Medium	Unvoiced
e (b**a**ke)	Front 2			Medium	Tensed
i (P**e**te)	Front 1			High	Tensed
o (b**oa**t)	Back 2			Medium	Tensed
u (l**u**te)	Back 2			High	Tensed
x (ab**ou**t)	Central 2			Medium	
A (b**i**te)	Central 1	Front 2		Medium	Tensed
E (s**e**t)	Front 1	Front 2		Medium	
I (b**i**t)	Front 1			High	
O (b**oy**)	Central 1	Central 2		Medium	Tensed
U (b**oo**k)	Back 1			High	
W (b**ou**t)	Back 1	Central 2	High	Medium	Tensed
Y (c**u**te)	Central 1	Front 1	Front 2	High	Tensed
@ (b**a**t)	Front 2			Low	
(**o**ne)	Central 1	Front 1	Glide	Low	Voiced
^ (b**u**t)	Central 1		Low		

present these relationships in alphabetical order. Tables 7.1 and 7.2 list the same data set but now grouped into vowels and consonants.

In Tables 7.1 and 7.2 the characters, presented to the input layer of the network, are given with their related phonemes, which are presented to the output layer. With every character the phonetic features that specify the pronunciation of the symbol are depicted. Articulatory features that are used to code vowels and consonants are shown underlined in the category in which they are used the least (e.g. 'Voiced' is used 21 times in coding a consonant and once in coding a vowel and is therefore underlined in the category 'vowels'). The tables show that this overlap is limited to 4 of the 51 symbols ('c', '*', 'X' and ':'). The tendency for vowels and consonants to be coded in nonoverlapping ways is violated only in the case of the letter coded as ':' (logic). The pronunciation of the consonant 'c' in this context, symbolized by ':', is coded with articulatory features that are mostly used for representing vowels. This, however, implies that for NETtalk ':' is a vowel. In all other cases there are other non-overlapping features available for explicitly defining a vowel as a vowel and a consonant as a consonant. These results indicate that the features that are used to code about 95% of the vowels only code about 5% of the consonants and vice versa. Only 8 of the 24 features show any overlap and are used for coding vowels *and* consonants. Notice, however, that this overlap is always rather limited. For instance, the feature 'Unvoiced' is used 12 times in encoding a consonant

Table 7.2 *Consonants: phonetic features that are used in coding both vowels and consonants are underlined in the category in which they appear least frequently*

Phoneme (example)	Articulatory				
b (**b**et)	Stop	Labial	Voiced		
d (**d**ebt)	Stop	Alveolar	Voiced		
f (**f**in)	Fricative	Labial	Unvoiced		
g (**g**uess)	Stop	Velar	Voiced		
h (**h**ead)	Glide	Glottal	Unvoiced		
k (**K**en)	Stop	Velar	Unvoiced		
l (**l**et)	Liquid	Dental	Voiced		
m (**m**et)	Nasal	Labial	Voiced		
n (**n**et)	Nasal	Alveolar	Voiced		
p (**p**et)	Stop	Labial	Unvoiced		
r (**r**ed)	Liquid	Palatal	Voiced		
s (**s**it)	Fricative	Alveolar	Unvoiced		
t (**t**est)	Stop	Alveolar	Unvoiced		
v (**v**est)	Fricative	Labial	Voiced		
w (**w**et)	Glide	Labial	Voiced		
y (**y**et)	Glide	Palatal	Voiced		
z (**z**oo)	Fricative	Alveolar	Voiced		
C (**ch**in)	Affricative	Alveolar	Unvoiced		
D (**th**is)	Fricative	Dental	Voiced		
G (si**ng**)	Nasal	Velar	Voiced		
J (**g**in)	Affricative	Alveolar	Voiced		
K (se**x**ual)	Affricative	Palatal	Unvoiced	Fricative	Alveolar
L (bott**le**)	Liquid	Alveolar	Voiced		
M (absy**m**)	Nasal	Dental	Voiced		
N (butto**n**)	Nasal	Palatal	Voiced		
Q (**qu**est)	Affricative	Labial	Voiced	Stop	Velar
R (b**ir**d)	Liquid	Velar	Voiced		
S (**sh**in)	Fricative	Palatal	Unvoiced		
T (**th**in)	Fricative	Dental	Unvoiced		
X (e**xc**ess)	Affricative	Central 1	Front 2	Unvoiced	
Z (lei**s**ure)	Fricative	Palatal	Voiced		
! (na**z**i)	Affricative	Dental	Labial	Unvoiced	
# (e**x**amine)	Affricative	Palatal	Velar	Voiced	
: (logi**c**)	Front 1	Front 2	High		

and only once in encoding a vowel. These results indicate that the emergent vowel–consonant distinction was already fully present in the input and output patterns presented to NETtalk. This surprising result forces us to re-evaluate the claims and interpretations made for this model.

NETtalk is put forward as a model that shows emergent symbolic

representations, demonstrating the power of subsymbolic computing. In the presented analysis it is shown, however, that the subsymbolic explanation of NETtalk's performance in pronouncing English words, expressed in the separation of vowels and consonants, is due to the encoding supplied by the designers of the system. The vowels are always translated to a set of articulatory features which themselves distinguish vowels from consonants. Therefore, it is not surprising that the hidden units of NETtalk learn to discriminate them. NETtalk just separates patterns that are sufficiently different and groups together patterns that are sufficiently similar. Since the categorization in vowels and consonants was already there in the patterns presented, it is no surprise that the system captures this regularity. The trick of this subsymbolic 'explanation' lies in the proper encoding of the desired symbolic behaviour into activation patterns that are presented to the network. This encoding is made by the designers of the system, Sejnowski and Rosenberg, and not by NETtalk (or its learning algorithm).

This result might sound trivial but the point is that the claimed emergence of symbolic behaviour from subsymbolic processing cannot be supported. Let us see how this applies to the three claims of Churchland and Sejnowski listed above. Claim 1 (the system finds the representations itself) is plainly wrong, as shown in the analysis of the patterns presented to the model and their relation to the precoded features. The representations it finds are completely prespecified in the examples it learns. These examples are defined by the designers. Claim 2 (the representations do not resemble sentence-logic organization) must be reformulated. It would be more appropriate to say that although the representations after learning are interpreted in terms of the activation of the hidden units, they are a direct result of the symbolic precoding made by the designers of the system (or rather the designers of DECtalk), and for that matter are completely related to a symbolic analysis of language production. Claim 3 (different initial conditions lead to the same functional clustering) can now be understood. It is obvious that, given different initial conditions, the system will always settle into the same functional clustering because it is forced into it by the presented target patterns (which represent the phonetic feature coding made by the designers). This set of patterns does not change over the different experiments. Given these results, the suggestion that NETtalk is somehow helpful in understanding brain dynamics is hard to interpret. This is emphasized by the serious criticism that the used learning method received for not being very brain-like (e.g. [Crick 1989]). Moreover, the claim of Sejnowski and Rosenberg that NETtalk did not start out with 'considerable' innate knowledge of the task domain cannot be supported. Not only was the network totally and unambiguously symbolically labelled at the level of input and output units, but the way these symbolic representations were engaged and associated was again explicitly coded in the set of patterns presented.

This analysis suggests that the subsymbolic strategy behind NETtalk consists of the following steps.

1. The designer of the system defines basic symbolic properties in which a certain task can be described (in NETtalk articulatory features and characters): the knowledge the system must have to accomplish the task is defined.
2. These properties get translated to regularities in activation patterns presented to a connectionist model (in the case of NETtalk this is expressed as which letter should be associated with which set of pronunciation features). The regularities expressed in the predefined input-to-output mapping encode the rules for the pronunciation of English text. The connectionist model learns to separate the patterns on their differences and groups them together on their regularities.
3. The rules, encoded in the weight distribution, are applied to input patterns presented to the network, which transforms them into the related output pattern. The rules encoded in the network lead to certain regularities in the dynamics of the network – the activation of the output and hidden layers – or a specific distribution of the weights.

These three steps are amazingly similar to the three characteristics of the cognitivistic approach distinguished by Winograd and Flores. An additional step, not identified by Winograd and Flores, is that the regularities expressed in the dynamics of the network are in turn symbolically interpreted by the designer (in the case of NETtalk as a vowel–consonant distinction). The most important commonality between the cognitivistic tradition and subsymbolic connectionism, as analysed here, is that both completely rely on designer-dependent symbolic task descriptions.

Of course, the analysis of the representations formed in models like NETtalk can be useful because 'unexpected' regularities (like the vowel–consonant distinction in NETtalk) can be discovered (see also [Rosenberg and Sejnowski 1987]).[3] We must not forget, however, that these discoveries only mean that the designers, who made the precoding, did not exactly know which regularities they put in. A more efficient way to discover them could have been to analyse these precodings directly instead of first making NETtalk learn them. (For instance, the quite straightforward rearranging of the data set presented in Table 7.1 is already quite insightful.) We have to be very clear about the claims based on the performance of NETtalk: the vowel–consonant separation is not an emergent property of the system. Moreover, compared to systems that use a set of letter-to-sound rules (without a dictionary of exceptions), NETtalk does fairly poorly [Klatt

[3] It must be said, however, that every introductory text on phonetics would show that the sounds of vowels and consonants are each described by a different set of phonetic features.

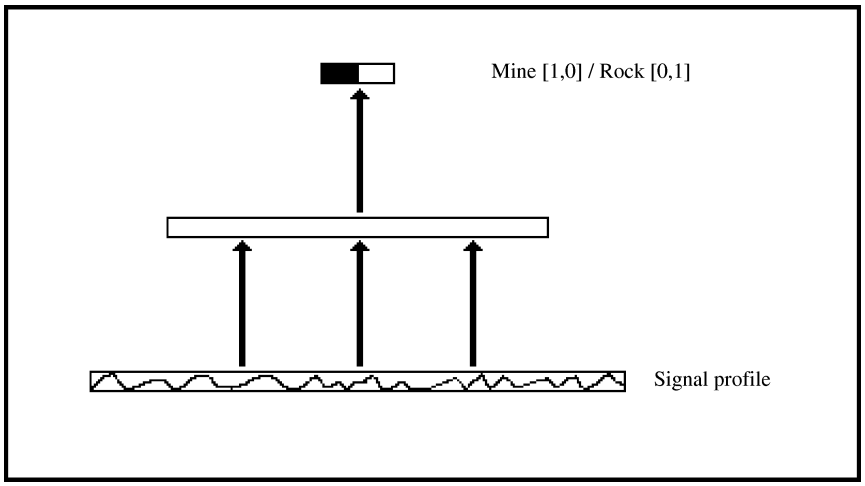

Figure 7.2 *The architecture of the sonar classifier: see text for explanation*

1987]. All it shows us is how we can 'compile' a given designer-dependent symbolic task description into a connectionist model.

To further explore the generality of this contention, two more examples will be considered. In the first example it will be shown that indeed no 'emergent' symbolic behaviour appears when the task domain is not symbolically defined *a priori*. The second example will show that the suggestions by [Churchland 1989] and [Ramsey *et al.* 1991], that subsymbolic reduction takes place at the 'lower' level of the weights or of the 'formal laws' driving a connectionist model are not supported by the nature of these systems.

7.6 Signal classification and subsymbolic reduction

[Gorman and Sejnowski 1988] have used a network with an architecture similar to NETtalk as a classifier of sonar signals returned by metal cylinders and rocks. They show that this task can be accomplished with an accuracy of 99.8% by a network that consists of 24 hidden units (see Fig. 7.2).

The patterns presented to the input layer correspond to the spectral information of the sonar return signals. The network is trained to associate these patterns with the corresponding two-bit pattern for cylinder [1,0] or rock [0,1]. At the input side the spectral information of an analog signal is presented and at the output side a symbolically interpretable discrete coding is employed.

Gorman and Sejnowski expected that, analogously to NETtalk, the global

features expressed in the activation patterns of the hidden units could be symbolically labelled. In this case the symbols 'rocks' and 'metal cylinders' seemed appropriate. Unfortunately this was not possible:

> 'Although it is attractive to think of a hidden unit as a feature extractor, this may not be the best way to characterize a hidden unit's coding strategy. As we demonstrated, the hidden unit is capable of encoding multiple features and even multiple strategies simultaneously. The network is able to internally encode pattern variations that do not decompose simply into a set of feature dimensions.' [op. cit. p. 88].

Eventually Gorman and Sejnowski resort to a description in terms of rules and strategies to illuminate the behaviour of their model. Although this can be very helpful in understanding the behaviour of such a complicated model, it implies that the regularities expressed in the analog signal projected to the input layer do not coincide with a (sub)symbolic interpretation of the task by a human observer. In this case symbolic behaviour did not 'emerge' because the task domain could not be defined symbolically. The regularities in the spectral information presented to the input layer that were captured by the model did not lead to activation patterns at the hidden layer that would allow a straightforward interpretation in terms of symbolic categories.

Subsymbolic reduction as proposed by Smolensky (and others, e.g. [Churchland and Sejnowski 1989]) seems to be closely tied to the possibility of completely describing the task in symbolic terms. The input to the sonar classifier was not precoded in discrete symbolic features. This input did not allow a mapping into the symbolic categorization coded by the output units. The interpretation of network behaviour in terms of symbols emerging from subsymbols became impossible: no symbols, no subsymbols.

The analysis of the sonar classifier suggests that subsymbolic reduction, as defined here, only seems to succeed if the patterns of activation of the model are interpretable in advance in terms of discrete symbols. Hence, the success (or failure) of subsymbolic reduction only indicates whether the designer succeeded in describing the task domain in advance in symbolic terms. This example further supports the previously listed set of rules employed by subsymbolic connectionists.

7.7 Donald Duck: the example

As already indicated, another subsymbolic route to understanding cognition is proposed by [Churchland 1989] and [Ramsey *et al.* 1991]. For instance, in [Ramsey *et al.* 1991] it is claimed that the coding of propositions in the weights of connectionist models is qualitatively different from a symbolic one. With an example it will be shown, however, that also at this level the same dependence on the symbolic description of the task by the designer holds.

Consider a connectionist model of Donald Duck who learns to represent his family relationships, such as his three nephews (the formal structure of this model is described in Appendix A). In the course of a day Donald Duck is confronted with Louie, Dewey and Huey. In this model I will use a model similar to one presented by [McClelland and Rumelhart 1986] describing a boy representing visual impressions of dogs and their names, in order to illustrate a way the three nephews get represented in Donald Duck's memory. The main idea of using this model is to argue against the claim implied by [Churchland 1989] and [Ramsey *et al.* 1991] that connectionist representations of this kind are qualitatively different in any way, and thus to issue a warning against an over-interpretation of such models.

Donald's learning is modelled with an autoassociative system consisting of eight units. Every time Donald is confronted with one of the patterns for 'Louie', 'Dewey' or 'Huey' they activate a specific pattern in these units. When Donald sees one of his nephews the patterns transduced to his memory by his sensors are:

```
[ +1, +1, +1, +1, -1, -1, -1, -1] = 'Louie'
[ +1, +1, -1, -1, +1, +1, -1, -1] = 'Dewey'
[ +1, -1, +1, -1, +1, -1, +1, -1] = 'Huey'
```

Donald is confronted with his nephews during 150 learning cycles. In every cycle first Louie presents himself to Donald, followed by Dewey and Huey. To evaluate the suggestion that the weights implementing Donald's memory somehow constitute a subsymbolic representation of his nephews, we must determine how these representations are coded in the weights. Evaluating only one weight (as proposed by [Ramsey *et al.* 1991]) is not appropriate.[4] The representations formed in the connections between the units can only be evaluated in terms of a configuration of weights. When we have a one-layered network an appropriate technique is eigenvalue decomposition (e.g. [Anderson and Murphy 1986]). This technique provides us with the patterns (eigenvectors) coded for by the weights and their prominence (eigenvalues). The results of the decomposition of the weight matrix of Donald's memory, in terms of eigenvalues and the covered percentages, are depicted in the following table.

value:	percentage:
5.2 10-4	76
1.3 10-4	19
3.3 10-5	5
1.1 10-20	0
3.3 10-18	0

[4] Although connectionist models are well known for their robustness, also called 'graceful degradation', pattern recognition using only one weight will become a problem with networks bigger than two units.

```
        2.5 10-22         0
        1.3 10-20         0
        7.4 10-19         0

total:  6.8 10-4        100
```

These eigenvalues suggest that the matrix constituting the memory of Donald has three dimensions. The eigenvector with the largest eigenvalue explains 76% of the variance of the configuration of weights. The second largest explains 19%. The third largest explains 5%. The other five eigenvectors each explain 0%. In the following table the values of the eigenvectors associated with the three non-zero eigenvalues (or the three dimensions making up Donald's mind) are depicted.

eigenvector element:	1	2	3
1	1.000	-0.996	0.996
2	0.999	-0.996	-0.996
3	-0.996	-0.999	0.996
4	-0.996	-0.996	-0.996
5	0.996	0.996	0.996
6	0.996	1.000	-0.996
7	-0.994	0.993	0.996
8	-0.999	0.996	-0.996

When we compare these eigenvectors with the patterns which code Louie, Dewey and Huey the resemblance is striking. The three dimensions making up the memory of Donald are exactly the patterns we have put in!

Also in this example subsymbolic explanation is completely dependent on the a priori symbolic description of the task made by the designer. The symbols or propositions the network has to learn are translated by the designer into patterns of activation conserving the category differences and similarities of the symbols. Symbolic distinctions are translated to distinctions between the patterns that code these symbols. By learning these patterns, the weights of the model will be adjusted to these distinctions. The subsymbolic representation encoded in the weights will be solely determined by the properties of the patterns the designer has put in. Thus the representation is *not* qualitatively different from symbolic representations, as [Churchland 1989] and [Ramsey *et al.* 1991] claim, but has the same dependency on the user's ontology.

7.8 The status of subsymbolic computing

The above three examples all support the hypothesis that subsymbolic connectionism, as defined and analysed here, is based on a circular strategy

culminating in a category error by the designers and interpreters of these models. The severity of this conclusion would make a further generalization necessary. It is, however, not the purpose of this chapter to provide a full review of the different realizations subsymbolic connectionism has found in cognitive science. The intention is to define a conceptual framework and research strategy in which the connectionist ambition can be realized. It is up to the designers of connectionist models, and any other models for that matter, to be critical towards their own strategy and interpretations. The designers of the models and their interpreters discussed here have failed to do this, leading to a further confusion about the scope and limits of the approach. A reasonable conclusion of the analysis presented would be that there is a tendency in the mainstream of connectionist modelling to over-interpret models that serve solely as compilers of observer-dependent ontologies. At the heart of this over-interpretation lies a category error where the ontology of the designer is mistaken for that of the model he or she defines.

The case against subsymbolic connectionism, as analysed here, seems convincing enough. Although some of these models can be considered engineering successes, their explanational value seems minimal. They only echo the symbolic regularities the designer has put in. Like the cognitivistic tradition, subsymbolic connectionism seems to start out with a symbolic analysis of the task domain. The regularities discovered in this analysis are translated by the designer to specific activation patterns presented to the model. This implies that the subsymbolic bridge between the symbolic mind and the dynamical brain is based on an unsubstantiated claim. The models brought forward as examples of its potential indicate that a subsymbolic analysis is completely dependent on a symbolic one. Both Smolensky and Newell, for instance, take symbols as patterns of activation in some hardware medium. The only difference between the two positions towards symbols is that the former sees an emergent relationship between these two levels.

A frequently offered interpretation of the status of connectionist models in cognitive science is that we do not know enough about these networks to come to a final evaluation (e.g. [McCloskey 1991]). The analysis presented, however, shows that we do know enough about these models to perform such an evaluation: the subsymbolic route to understanding cognition, as defined and analysed here, is strongly dependent on the well-trodden symbolic one.

A standard reply to this type of criticism is that it might be true for these 'old' examples, but in the meantime much has changed. It is true that the last few years have seen an explosive growth in the knowledge of the techniques and methods of connectionist modelling. In the case of these more principled issues, however, not much progress has been made. For example, in a recent popularization of connectionist modelling as applied

to language and language disorders due to brain lesion, [Hinton *et al.* 1993] (see also [Hinton *et al.* 1991]) present a model that is designed following the same strategy as NETtalk.

Let me emphasize again that the purpose of this analysis is not to provide a full review of the past and present use of connectionist models in cognitive science. The analysis of the traditional cognitivistic paradigm served to identify a number of conceptual issues relevant to solving the mind–brain dilemma. The evaluation of subsymbolic connectionism showed that this style of modelling does not automatically solve these problems, but in fact runs the risk of importing them by implicitly following the cognitivistic explanational scheme. It cannot be excluded that subsymbolic connectionism can provide a way to settle the mind–brain dilemma. This promise, however, still awaits its realization.

7.9 Taking connectionism seriously

The properties that render certain models connectionist are by reference to parallel computation, distributed representation and emergence. They comprise a set of techniques that can be used to describe dynamical systems. These techniques, however, can be applied to model a number of different biological phenomena like the brain, the immune system, co-evolution, and auto-catalysis (for a comparison see [Farmer 1990]) to mention but a few. The school of connectionism in cognitive science proposes to use these techniques to study the brain and behaviour. By itself, however, using these techniques does not constitute a paradigm. They are neutral towards a conceptual interpretation. Therefore, their relevance to cognitive science depends on the theoretical context in which they are embedded. It seems that in the current use of connectionist techniques three global domains can be distinguished: symbolic models, neutral models and neural models.

The first class of applications of connectionist techniques, symbolic models, combines connectionist techniques with a cognitivistic theory expressed in a symbolic description of a task (derived using step 1 of the cognitivistic approach). The models proposed in the 'subsymbolic school' (as defined and described earlier) fall into this category. As demonstrated earlier, in NETtalk the symbolic analysis is expressed in the letter and phoneme relations which are precoded in the patterns the system has to learn. This strategy of connectionist modelling must be considered as the application of a connectionist methodology within the traditional cognitivistic paradigm. Examples of this popular category of models are NETtalk, TRACE, the past tense model of [McClelland and Rumelhart 1986], a recently proposed model by [Mitchell and Hofstadter 1990] which demonstrates the 'emergence' of understanding, the models of [Seidenberg and McClelland 1989], and of [Hinton *et al.* 1991] on language, and the model on categorical and spatial representations presented by [Kosslyn *et al.* 1992]. A confusing

aspect of this type of connectionist models, as the analysis presented has shown, is that they are supposed to be something other than they are. This ontological fallacy is, unfortunately, not widely acknowledged. It is, for instance, only in the analysis presented that the ontological commitment and the semantic interpretability of this type of connectionist models is made explicit.

The conceptual dependence of these models on cognitivism implies that they have to deal with the standards set in that paradigm. They must answer the criticism from the cognitivist tradition (like [Fodor and Pylyshyn 1988] and [Broadbent 1985]), for instance that the representations built in connectionist models must satisfy constraints such as compositionality. Moreover, the analysis presented shows that the contribution of these models to our understanding of mental phenomena is limited since they only echo what their designers had already expressed in their symbolic task definition. Hence, they can never lead to an understanding which goes beyond the limits set by the cognitivist paradigm and its symbolic description of behavioural regularities.

The second domain of the application of connectionist techniques, neutral models, strictly limits their scope to the study and application of the techniques, independent of the question whether these models relate to cognitive processes or neural mechanisms. In contrast to the first class of models, in this case the units are *not* externally labelled with symbolic information, but are taken as neutral processing elements. In this perspective one can work on, for instance, convergence proofs and analytical methods to study network behaviour. Examples of this approach are the evaluation of the perceptron by Minsky and Papert, work on connectionist learning rules (e.g. [Rumelhart *et al.* 1986] and [Ackley *et al.* 1985]), the relation to techniques stemming from statistical physics (e.g. [Amit 1989]), or the work on the relationship between these connectionist techniques and the concept of universal approximation, which stems from the mathematical domain of approximation theory [Hornik *et al.* 1989]. This class of connectionist efforts can be seen as further improving and understanding connectionist techniques. This understanding can be used to address engineering problems like classification. It is an effort independent of the way in which these techniques could be applied in the domains of neuroscience or cognitivism, and seems more to rely on formal analytical methods. Like in the previous class of models the brain might serve as a metaphor in modelling, but it is not seen as a constraint.

A third and last class of connectionist models, neural models, relates to the application and development of these techniques in the domain of neuroscience. Classical examples here are, for instance, the work of [Hodgkin and Huxley 1952] on modelling the properties of single neurons, or the model proposed by [Willshaw and Malsburg 1976] on the formation of topological maps. An important and more recent example of this type of research is

the approach of synthetic neural modelling proposed by [Edelman 1987], [Edelman 1989] and [Edelman 1992], who tries to explain psychological phenomena from a (neuro)biological perspective using neural models. This class of models distinguishes itself from the previous two since it is not only inspired by the properties of the brain, but is also explicitly validated against strict constraints stemming from it.

It is not always possible to draw clear boundaries between these three classes of connectionist modelling. We have to keep in mind, however, that the techniques used to construct a model are neutral towards their interpretation. A model does not fall in the category of neural modelling just because basic processing units are called 'neurons' and their interconnections are called 'synapses'. They only become interpretable when the conceptual framework in which they are embedded is explicated and the models are validated against sufficient constraints. Sections 7.10 to 7.15 will be devoted to defining such a framework and elaborating the issue of constraints in the context of the initial connectionist ambition of solving the mind–brain dilemma: to understand behaviour and the mind from the perspective of the third domain of application of connectionist techniques, namely neural models. The proposal will be split into two aspects. The first relates to the methodological issues involved, while the second focuses on the conceptual ones. This proposal will be further defined by discussing a concrete model that is developed according to the principles put forward.

7.10 Constraining connectionism: a research strategy

[Massaro 1988] showed that a connectionist system (in particular the widely used generalized delta procedure which is also employed in NETtalk) can be used to model mutually exclusive psychological models of perception. This *superpower* of connectionist models is understandable when we realize that the most important connectionist learning mechanisms are implementations of well-known optimization techniques (like gradient descent and simulated annealing). For instance, [Wray and Green 1991] have shown that these modelling techniques are equivalent to traditional approximation methods like polynomial approximation and Volterra series. Moreover, [Hornik *et al.* 1989] proved that any well-defined input–output mapping can be approximated by a multilayered connectionist network. Hence, being able to train a network to perform a certain task does not ensure that the network is a psychologically plausible model of this task. It shows that one has been able to *transform* the task to a well-defined set of input–output mappings that a powerful optimizer could learn. This implies that when we want to use connectionist techniques to explain psychological phenomena, additional constraints must be met.

Given the neutrality of connectionist techniques and the superpower of some of them it must be shown that they are not only capable of learning

a given input–output mapping, but also that they perform in a realistic way. Until now the opposite seems mostly to have been the case. It is disappointing that insofar as connectionists succeed in getting their models to run, practically none of them satisfy any relevant psychological or (neuro)biological constraints. Examples of this can be found in the criticisms raised by Massaro against popular connectionist models of speech perception based on interactive activation (e.g. [Massaro 1989]), the critical analysis offered by Pinker and Prince of connectionist language models [Pinker and Prince 1988], or the already mentioned critique by Crick on the biological plausibility of the standard practice of connectionist modelling [Crick 1989].

The choice of constraints is, of course, closely related to the choice of phenomena to be studied. Subsymbolic connectionism immediately tried to deal with sophisticated psychological processes such as language (which have also been the focus of cognitivism). Above I tried to show that these attempts did not explain very much beyond what was already known (or could be known) by sticking to an exclusive symbolic analysis. This, of course, raises the question whether connectionism has been focusing on issues it can deal with and whether it has followed an appropriate strategy.

I argue that the selection of constraints to be met by a connectionist model is not an arbitrary process. The minimum standard that we should impose on this process is that constraints should be drawn from *both* the domain of behaviour and the level of implementation, neurobiology. This standard may sound rather obvious, but this simple rule does not seem to be followed by the mainstream of connectionist modelling. As, for instance, indicated by [Uttal 1990], the functional descriptions of behavioural regularities provided by information processing models are in principle underdetermined. This means that they can never provide sufficient constraints on (connectionist) models that try to account for these behavioural regularities. This argument goes back to the work of [Moore 1956] who showed that it is in principle impossible to decide between alternative functional models of an observed response function. Behavioural constraints, therefore, cannot be taken as sufficient guidelines but must themselves be validated. Therefore, aside from finding constraints at the behavioural level, it is necessary to add constraint from another level: the brain. In a more general sense one could say that in modelling a certain response function, whether it is one of a behaving organism or of a firing neuron, the level at which this response function is observed does not suffice as a source of constraints. Constraints pertaining to underlying mechanisms need to be included.

It should be added here that Uttal believes that the brain cannot provide constraints on models of behaviour since the brain is too complicated and highly non-linear. Uttal is correct with this latter observation, but that does not imply that relating psychological phenomena to brain mech-

anisms is *in principle* impossible. These properties of brain dynamics indicate that a straightforward decomposition of those brain processes that implement or give rise to psychological phenomena might not be possible. We should not forget, however, that the picture of brain mechanisms sketched in terms of nonlinear dynamics might not be very compatible with the information-processing framework of psychological models. This relatively recent development in natural science can also be taken as an indication that psychology should change its conception of the processes and mechanisms underlying the phenomena it studies. The conceptual issue of the mind–brain dilemma can be taken as an additional indication of the incompatibility of these levels of description.

Results from neurobiology, however, cannot be the only source of inspiration for connectionist modelling. The data generated by this field will certainly get us lost in the many details of neural functioning. It would be naive to believe that we can automatically solve all sorts of issues in the realm of the behavioural sciences by just turning towards neuroscience. Moreover, the data pertaining to certain brain structures or mechanisms can be quite confusing. For instance, in the debate on the properties of dopamine receptors in the brains of schizophrenic patients one research group reported elevated sensitivity of these receptors [Wong *et al.* 1986] while others found evidence for normal sensitivity [Farde *et al.* 1987].

Does this mean that Uttal was right after all? I would claim this is not the case. Both the study of behaviour and cognition and the inquiry into the neural substrate are developing fields, where we do not even know whether we are posing the right questions. Both these levels of description – in fact the multitude of levels at which behaviour gets expressed and generated – need to be related to each other to address the indeterminacy from which they suffer in isolation. Behavioural, top-down, constraints are necessary to give guidance in picking the properties of brain dynamics that are crucial in understanding its functioning (see also [Clark 1989]).

Only by cross-validating connectionist models against behavioural *and* neurobiological constraints can we expect these models to be able to contribute to our understanding of psychological phenomena. This strategy of convergent validation allows an ongoing and necessary interaction between the behavioural and brain sciences. First, constraints from both domains are integrated in a model. Next, the results of this integration process will lead to new hypotheses that can be communicated back to the involved levels of description and tested in these domains. This can lead to a further specification of the constraints applied to the model. Therefore the role of modelling becomes one of integrating multiple levels of description and finding convergence between them rather than mimicking the regularities found at one of them.

7.11 From knowledge to adaptation: a conceptual framework

A research strategy will not automatically lead to the understanding of a phenomenon if it is not embedded in a proper conceptual framework. A conceptual framework can only be called 'proper' when it is explicated. Subsymbolic connectionism has focused on a symbolic characterization of behaviour and the knowledge underlying it. The behaviour was explained on the basis of assumed internal symbolic representations. The same is true for traditional cognitivism. [Newell 1990], for instance, indicates that the knowledge level does not explain the aboutness of representations but assumes it. This can be a useful position, but still cannot, by itself, be considered a firm base for theorizing (see also [Smith 1990]). Knowledge itself, rather, seems to be a phenomenon that must be explained. [Harnad 1990] has pointed out the problem of symbol grounding: from what do symbols get their meaning? The standard strategy of cognitivism is to assume that symbols derive their meaning out of their relation to other symbols, like the input patterns to NETtalk derive their meaning from their predefined association with the pronunciation features. In this way we end up in a closed loop of interrelated symbols. In the case of NETtalk this boils down to answering the questions as to how the pronunciation features got there in the first place and why a specific input pattern should be associated with a specific output pattern.

There are two approaches to explain the grounding of the knowledge of a system. The first one, implied by cognitivism, is putting the responsibility on evolution and assuming that rules and representations are just handed over to a system by its genetic code [Piatelli-Palmarini 1980]. This brings us back to the problems of nativism discussed earlier. A more realistic solution supplements genetic predefinition with learning. A system acquires knowledge by interacting with its environment. This process is guided by the genetically defined set-up of the system, its phenotype, and the properties of its environment. [Bechtel 1989] indicates that connectionism promises to explain the aboutness of representational states. It is exactly this promise, I think, that connectionism should focus on. It might be useful to explore where knowledge comes from, and how it is defined in the adaptive structures that we call organism and brain, and around which primitives it is organized and expressed. I would like to emphasize that the nature of knowledge should be explored before elaborating on its use. There is no claim of originality here. This proposal is very much along the same lines as, for instance, the school of genetic epistemology of [Piaget 1971] (for an overview see [Furth 1969]).

In concentrating on the capacity of connectionist systems to 'learn' by changing their structural properties, the strength of their interconnections, they provide a set of techniques which can be used in addressing the symbol-grounding problem. They can be used to study adaptive structures, like

the brain. We should, however, not make these systems learn our symbolic descriptions of specific tasks, but concentrate on how they can acquire the behaviours necessary to perform these tasks out of their interaction with the real world. This interaction is mediated by the phenotype in which these structures are embedded: the morphology of the body, and the properties of the sensors and effectors. This implies that, on one hand, symbols cannot be seen as part of the internal mechanism that determines behaviour. This internal mechanism must be seen as a control structure. It controls since it mediates and transforms sensing into acting. It is a structure since it is realized in the physical world. There is no reason to assume that the general principles that generate this structure and that are implemented by it will coincide with the formal logical picture of the mind put forward by cognitivism. Rules and representations should, therefore, be taken as observer-dependent constructs that are used to describe the behavioural regularities generated by the behaving system. On the other hand, the learning methods used cannot be supervised like backpropagation. This is not only because of its biological implausibility, but more specifically for its tendency to allow a human designer to compile their own ontology into a connectionist model. It might lead to acceptable engineering, but most definitely not to any insight into the functioning of the brain. The brain can only be its own teacher.

A standard argument against this position is that by using supervised learning methods one can find an existence proof of a possible neural structure implementing a certain function. [Churchland 1989] argues that as long as one shows that the architecture is sufficiently similar to neural anatomy and the dynamics modelled capture the general properties of neural dynamics, then the use of a supervised learning method can be justified to generate ideas on ways in which the brain could be configured to perform certain tasks. Given the previous analysis we have to conclude that in the standard examples of the subsymbolic paradigm the designers have not demonstrated that their models satisfied any relevant neural constraint. The main source leading to their design decisions seems to have been their own functional task decomposition. This raises the question as to why there is so little biological evidence supporting these models. The present analysis suggests that the solution to this problem has to be traced back to the ontological errors the designers have made in the design of their systems. By assuming that the world presented itself in a prediscretized way, and that learning equalled direct supervision extending to every weight in the system, the designers could but end up in a realm which was beyond biological relevance. It is not so much the case that the proposed models were constructed to solve a problem in a different way than the brain might do it, but that they basically tried to solve a totally different problem. The first problem a brain has to solve, and so for any theory explaining it, is how to categorize the world (see also [Edelman 1987]).

As argued before, the symbol-grounding problem, as defined in the strict sense, can be interpreted as an artifact of a symbolic approach. Moreover, the issue of levels and the problem of complete reduction acquire a different status in the present proposal. As described earlier symbol manipulation is traditionally taken as a separate level of explanation that can be related to a level of neural implementation. Subsymbolic connectionism tried to establish this relationship. In the present proposal, however, this type of relation might turn out not to be the one to look for. Before expanding on this claim I will discuss a concrete example which illustrates the previously identified methodological and conceptual issues.

7.12 Interacting with the real world

This proposal is based on several assumptions relating to the nature of the interaction between an agent and its environment.

An agent is defined as the conglomerate of its phenotype and its control structure, both expressed as physical structures. I prefer to refer to the nervous system of the agent as a control structure to emphasize that in this perspective a nervous system is modelled and not literally copied. The principles expressed in the control structure must be seen as a hypothesis regarding the principles implemented in real nervous systems. At this moment in the history of cognitive science it might be useful not to overstate our claims. The environment of the agent, the world, is defined according to the following assumptions.

1. The real world is only partially knowable and only partially predictable.
2. The world does not consist of a collection of discrete events.
3. The world has its own temporal dynamics.

These three assumptions have important implications for the agent. Assumption 1 implies that there cannot be a predefined body of knowledge that captures the pertinent properties of the real world. Assumption 2 forces us to conceive the input to the system, its sensory states, as continuously varying and not as discrete. The notion 'event' is completely connected to the continuous interaction between a system and the world. In fact, categorization is the active creation of an event. The last assumption indicates that the agent is under time pressure to act. This imposes a severe constraint on the mechanisms that mediate and transform sensing to acting.

The methodological implication of these assumptions is that if we want to find a way to deal with the mind–brain dilemma, which is the ambition of the connectionist programme, connectionist models should be applied in the context of autonomous agents. This implies that the models, or control structures, are embodied in artifacts that have sensors and effectors and that interact with the real world. They should be complete models that

span the whole domain from sensing to acting. Moreover, they should be in a constant interaction with the environment, and be able to survive in it.

Moreover, given these assumptions we have to conceive of autonomous agents as adaptive systems. This raises the question *when* and *why* a system should learn. Assumption 1 answers the why question. This does not mean that *nothing* can be foreseen in advance. For instance, what the food looks like that a specific animal feeds on can be predicted; where it can be found in its environment, however, cannot be predefined. The empirical fact that the genetic code has only a limited coding capacity and can therefore not be expected to code the complete body of knowledge might therefore imply that instead of being 'short on memory' the genetic code only predefines what can be predefined and will leave dealing with uncertainty in the real world to the mechanisms for adaptation and learning.

This proposal is not entirely new. Its synthetic aspects, for instance, go back to the work of Hull, who in the 1920s was already trying to develop a mechanism that could learn according to the principles of classical conditioning [Hull and Baernstein 1929]. Another relation can be drawn to the emerging field of 'new AI' [Brooks 1991a, Brooks 1991b]. Contrary to the former the proposal not only focuses on mechanistic models but also on a specific theoretical framework where grounding and situatedness are key issues. The contrast with the latter is that 'new AI' seems to draw its main inspiration from engineering and intuitions about cognition. Although the intuitions relating to situatedness and embodiment seem reasonable they have to be placed on a more explicit base. Also, in fact, the paradigmatic examples of the 'new AI' movement suffer from the same sort of ontological problems as subsymbolic connectionism (see [Verschure *et al.* 1992] and [Verschure 1992] for a further analysis). In its realization the proposed framework shows a strong similarity to the work of [Edelman 1987], [Edelman 1989] and [Edelman 1992]. The main contrast is that Edelman places a strong emphasis on the biological components of this research programme, while the present proposal is defined against a more conceptual analysis of alternative approaches (see [Verschure and Pfeifer 1993] for a further analysis). Moreover, Edelman contrasts his proposal with an empiricistic one, where one assumes that an adaptive system is instructed by the world. The present proposal is derived from an analysis of the rationalistic perspective behind cognitivism.

7.13 Distributed adaptive control

In [Verschure and Coolen 1991] a connectionist model of classical conditioning was proposed which had been developed in accordance with the research strategy defined earlier. Classical conditioning is a phenomenon that is very well suited as a starting point for developing autonomous agents according

to the demands laid out above. It is one of the basic learning mechanisms many animals have at their disposal to adapt to their environment by forming associations between sensory states. It is a domain that provides a large amount of empirical data that allows the definition of sufficient constraints on models that are supposed to explain it.

The behavioural constraints against which this model was validated were the Rescorla and Wagner laws of classical conditioning [Rescorla and Wagner 1972]. The neurobiological constraints the model had to satisfy were no pre-wiring of associations between stimuli and response (as opposed to paradigmatic examples from the field of reinforcement learning, e.g. [Sutton and Barto 1981]), which brings us back to the ontological issues discussed earlier), changes in plasticity are based on local unsupervised mechanisms, and function arises out of the dynamics of populations of units. We showed that these sets of constraints could be successfully brought together in a connectionist structure. To incorporate this model of classical conditioning into the conceptual framework described earlier a control structure for an autonomous robot, distributed adaptive control (DAC), was proposed [Verschure *et al.* 1992]. Let us focus here on the principles behind this control structure and how the model pertains to the conceptual issues under consideration.

The basic idea behind this control structure is that an adaptively behaving system moves from a stage of coarse adaptation, which is expressed in simple genetically predefined reflexes, to a stage of fine-tuned adaptation as a result of its interaction with the environment. This interaction affects the behaviour of the agent through learning. The reflexes are expressed in relations between primitive sensors (in this example a collision detector and a target sensor), which in general are taken to be proximity sensors, and simple motor programs (in this case avoid and approach actions respectively). Moreover, the structures involved in these reflexive actions form the criteria for learning in the system. These basic properties can be viewed as a value scheme [Edelman 1987], which can be seen as defined by the genetic set-up of the system: the 'genetic envelope' [Changeux 1985]. The value scheme defines the set of unconditioned stimuli and unconditioned responses, and the basic properties of the sensory and motor systems. In addition to these properties, the value scheme contains the mechanism that allows it to integrate its distal sensors into these actions. Fine-tuned adaptation is expressed in the engaging of the distal sensors (in this case a range finder) in these reflexive sense–act relations. This latter stage of behaviour is under the control of the distal sensor and the history of the system. This is expressed in the way the distal sensors are connected to the motor programs and the way the motor programs have changed over time (this last property will not be dealt with in the present illustration). Thus the control of behaviour over time shifts from the environment, through proximal sensing, to the organism, through distal sensing.

This approach immediately provides an alternative perspective on one of the fundamental controversies in psychology: whether perception is direct or indirect. Both stances to this problem seem to capture a part of behavioural reality, since they both can be justified on the basis of experimental data. When we, however, include the dynamics of epigenesis, as for instance expressed in DAC, it becomes clear that these viewpoints are two extreme positions on the developmental scale. This would also fit with well-known neurophysiological data. For instance, [Hubel and Wiesel 1962, Hubel and Wiesel 1968] showed that the ability to acquire certain perceptual categories depends on the presence of the right stimuli in the right critical period. Later on in epigenesis these perceptual categories increasingly control the interaction with the world. This would suggest that epigenesis moves from a stage of direct perception, in which the sensational invariants of the system–environment interaction get expressed in the perceptual mechanisms of the system, to a stage of indirect perception, in which internalized environment-related categories structure perception. Hence, the 'knowledge'-driven aspects of behaviour are understandable in terms of an evolving adaptive system. This implies that 'knowledge' underlying the behaviour of an organism cannot be analysed disconnected from its history.

In Figure 7.3 the phenotype of an agent used in the initial set of experiments with DAC is depicted. This agent has to deal with a (target) approach – (obstacle) avoidance task.

The left and right front sides of the system function as collision sensors. They will become active when the system touches an object at these locations. Colliding constitutes an unconditioned stimulus which is mapped onto a group of units. Activation in this group will lead to a motor output consisting of a retract-and-turn motion (9°) in the opposite direction to the collision. Because of its relation to an avoidance response, this group is referred to as 'the negative unconditioned stimulus' group (US–).

The target detector of the system, which could be interpreted as consisting of two 'ears', is sensitive to the difference in intensity detected by these two sensors. The unconditioned response will be to turn into the direction where the highest intensity is sensed. The group representing this unconditioned stimulus will be referred to as the 'positive unconditioned stimulus' group (US+) since it relates to approach actions. When approach or avoidance are not activated, the default behaviour of advancing (which can be seen as a simple form of exploration) will be executed.

The distal sensor, which senses the conditioned stimulus (CS), is a range finder which gives an inverse distance measure. This sensor covers a region in between -90 and 90 degrees from the front of the system and consists of 37 elements which each have their own receptive field. These receptive fields do not overlap. The angular resolution of these receptive fields decreases as the element is placed closer to the centre of the agent. Every element of

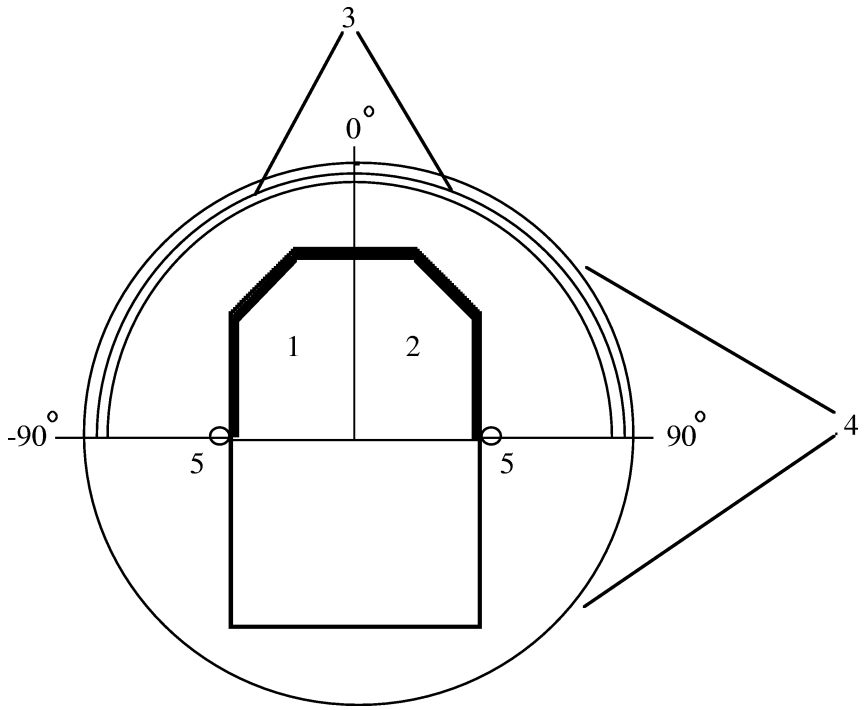

Figure 7.3 *The agent with its sensors. 1 and 2: left and right collision sensors; 3: region covered by the range finder; 4: the receptive field of the target sensor; 5: location of target sensors*

the range finder projects onto one element in the CS group. The activation of every unit of the CS group is proportional to the inverse distance in the receptive field of the range finder element that projects to it. This group of units can be seen as responding to time to contact [Lee 1976].

The group that codes the motor programs of the system, the unconditioned responses, consists of a number of so-called *command neurons* which code the motor responses [Kupferman and Weiss 1978]. This group will be referred to as the 'unconditioned response' group (UR). Whenever one of these command neurons is activated, a specific motor response is automatically executed. The connections between the US groups and the UR group are pre-wired and not modifiable.

The global layout of the control structure is depicted in Figure 7.4.

In addition to the already described groups that make up the anatomy of the control structure, a specific relation is defined between the approach and avoidance groups. Since it is more important for the system to avoid nearby obstacles than to approach targets (this is analogous to the conflict

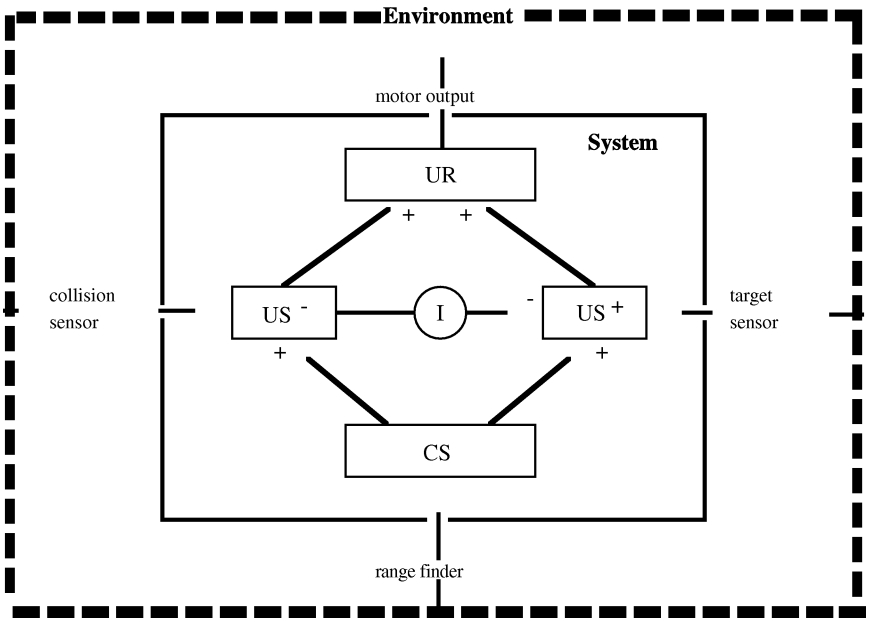

Figure 7.4 *The control structure and its relation to the environment. See text for explanation.*

theory by [Miller 1959]), activation in US– will inhibit the output of US+. In this case, activity in US+ cannot trigger approach actions. This inhibition only slowly decays, which means that US+ will be inhibited for a relatively long period of time.

In [Verschure *et al.* 1992] it is shown that this system can successfully learn to avoid obstacles and find targets. This means that the expected transfer of a reflexive avoidance or approach response, which is triggered by one of the proximity sensors, to a learned one, which is triggered by the distal sensor, has taken place (Chapter 5 by Pfeifer and Verschure gives a short overview of further explorations with distributed adaptive control, such as its robustness in parameter space, tests with different phenotypes, its 'psychophysics' and its performance with different sensors). In this chapter I will limit myself to its relevance to the conceptual issues. Before expanding on these examples I would like to emphasize some of the aspects of the present proposal and its realization in DAC. It is important to understand that the experiments described here are replicated in different simulation environments and hardware platforms, and do not represent a chance effect. They not only illustrate a robust way to achieve sensori-motor integration, but especially emphasize that an analysis of the methodological and conceptual issues raised can lead to successful modelling. Although DAC is

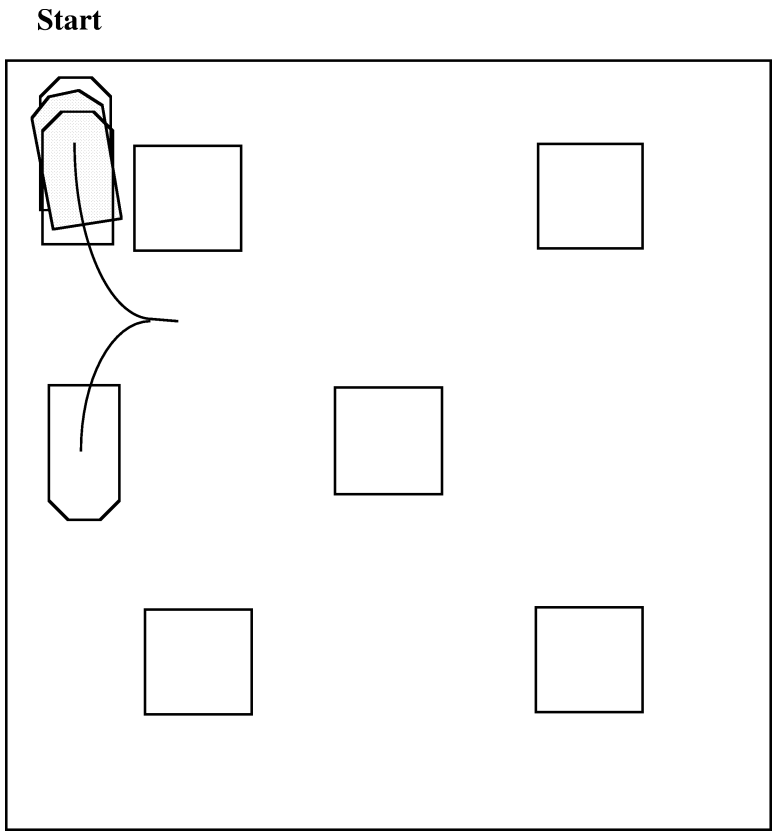

Figure 7.5 *Pulling out of an impasse. See text for an explanation.*

only a modest first step it clearly shows that the methodology and conceptual framework behind it are not defined with reference to a promise that awaits its realization in the future. This promise is realized. Before elaborating on this issue the aspects of the behaviour displayed by the agent that are pertinent to the present discussion will be discussed.

When the system is trained to avoid obstacles, and is then put in a difficult situation which it has never encountered before, it finds the solution shown in Figure 7.5. (For illustrative purposes the main components of this behaviour are redrawn – see [Verschure *et al.* 1992] for the original data.)

The agent starts out between an obstacle (the block at the right-hand side of the figure) and a wall (this location is indicated with 'Start'). The system subsequently backs out until it can make a complete turn. The symbolic description of this behaviour would be that a strategy to back out of difficult situations or for impasse resolution is executed. This is also precisely

the way in which standard robot architectures deal with these types of problems: the execution of predefined strategies programmed in by the designer (see [Malcolm *et al.* 1989] for an overview of traditional and more recent approaches in robotics). In this case, however, this behaviour has to be explained at a more primitive level. What looks like a well-organized behavioural pattern from a macroscopic point of view is in fact a sequence of local 'decisions'. The first step of the system is to retract and turn to the left (indicated by the dotted icons). This brings the system closer to the wall. The state of the range finder triggers further avoidance responses: retract and turn to the right and so on until the impasse is left behind. These responses trigger each other through the environment. Hence, behaviour at moment t is only dependent on the sensory state produced by the situation in which the system ended up after the motion made at moment $t-1$. The actual relation between this sensory state and the response made is defined in the connections between the different neural groups (CS, US– and US+) which express the invariants in the learning history of the agent. This chain reaction of local decisions or reactions leads to a seemingly well-organized behavioural pattern which can be interpreted by an observer as being determined by a behavioural strategy. This inferred strategy, however, has to be seen as an emergent property of the interaction between the phenotype, its control structure, and the environment.

Another example of this emergent relation between inferable behavioural strategies and the underlying dynamics of system–environment interaction is depicted in Figure 7.6. After having learned to avoid obstacles while trying to find a target (indicated by a black dot in the picture), the attractive force of the target was removed from the environment (the system could no longer 'hear' the target). When in this case the agent started to move around from its initial position (again indicated by 'Start') it found its target in a minimal number of steps. The most surprising behaviour, however, was that the system had learned to follow a wall. Again, the standard approach to making autonomous agents follow walls is just to predefine a specific behavioural strategy for it [Malcolm *et al.* 1989, Beer 1990]. Also in this case, however, the behaviour emerged out of the dynamics of the interaction between the agent and its environment. (A comparison between this notion of emergence and that employed in the subsymbolic paradigm will be made in the next section.) During its initial experiences in this environment, while the attractive force of the target was present, the agent always approached the hole in the wall, which hid the target, parallel to either the lower or upper wall. It had ended up in this position due to the sequence of avoidance movements (see [Verschure *et al.* 1992] for a complete description of this task). The environment was set up in such a way that the system could only detect the target when it was close to the hole in the wall. This implies that only in these cases was the sensory state

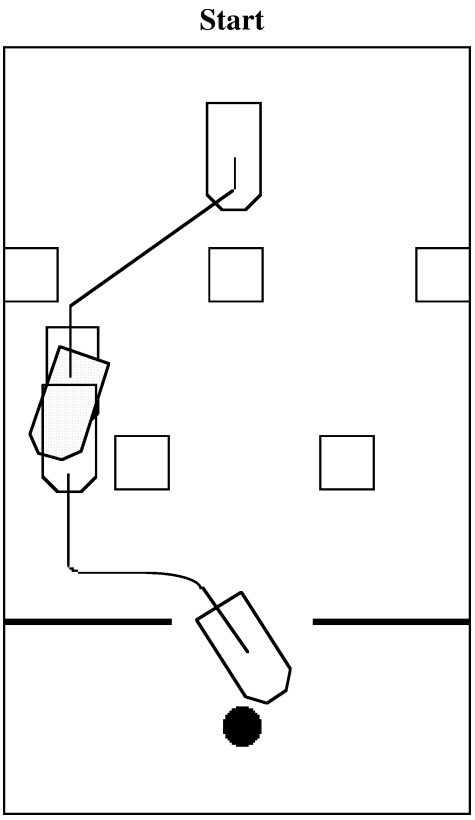

Figure 7.6 *Wall following. See text for an explanation.*

of the range finder associated with approach actions.[5] The representations that were formed in this way represent sense–act relationships that can be described as 'when you move parallel to a wall, turn towards it'. As soon as the agent found itself parallel to a wall it would turn towards it. This in turn would trigger the avoidance group since the agent had also learned to avoid obstacles. What looks like a coherent behavioural strategy is again a chain of local reactions: a sequence of approach and avoid actions. When the agent, after following a wall in this way, found itself in a hole in the wall the avoid group would not be triggered any more. The agent would pass through the opening and locate the target. This behavioural strategy was not predefined in the system. Following walls was not a task that was

[5] This decision was made in order not to cheat. When the attractive force of the target would have been equally strong in the whole environment it would in all cases have pulled the system towards it. This would have rendered any 'learned' approach behaviour not very convincing.

being modelled here. The control structure in no way supports the processing of sequences. Also, this behaviour can only be understood in terms of the interaction of the agent with a specific environment over time.

The full explanation of this behaviour would have the following form. At its starting position the system first avoided the obstacle in front of it and turned to the right. This brought it near another obstacle, which it avoided by turning to the left. Now the system had placed itself parallel to a wall. The following approach motion was triggered by the learned 'approach parallel walls' representation and a turn to the right is made (indicated with the dotted icon). This behaviour activated an avoidance response (the system turned left again). As indicated, activation in US– (which leads to avoidance movements) will inhibit the possibility of US+ (the approach responses) to trigger an action. Until this inhibition had decreased sufficiently to allow US+ to influence the actions of the agent, it proceeded in a straight trajectory. The next approach action could only be triggered when the inhibition of US+ dropped away. This made the sequence repeat itself. Again in this case, behaviour that might be very understandable in the macroscopic vocabulary of rules, strategies and representations can be generated by microscopic dynamical mechanisms which do not mimic these descriptions.

Of course these examples are disarmingly simple but they illustrate one important point: although behaviour might look very understandable in symbolic terms the structures generating it might do something completely different. Now the question must be addressed as to how this approach towards understanding behaviour bears upon the central theme of the present discussion: the mind–brain dilemma.

7.14 Comparing NETtalk and DAC

One could claim that like NETtalk, which was used as representing the basic thrust of the subsymbolic approach, the discussed control structure shows no emergent properties, since in this case the emergent behaviour is also a direct result of the way in which the designer has defined the system and its environment. I would like to show that this assertion is incorrect and that there are qualitative differences between the two models.

In NETtalk the 'environment' consisted of a set of letters which had to be mapped into the phonetic features encoded in the model. This 'environment' was not a physical environment accessed by the system through its sensors and influenced by its effectors, but a prediscretized set of input and output vectors. This training set, provided by the designers, defines which input vector should be related to which output vector. By analysing the predefined mapping of NETtalk it was shown that the 'emergent' properties of the model were contained in this predefinition. The vowel–consonant distinction was already unambiguously present in the precodings. Therefore,

the 'emergent' vowel–consonant distinction was completely predictable, given these precodings. The structure was defined according to the designer's domain ontology and learning proceeded by mirroring the designers domain ontology expressed in the input and output vectors.

The autonomous agent described had to adjust to the regularities present in its environment of obstacles and target, guided by its basic reflexes. The world was accessed through the sensors of the system. Its perception was influenced by its actions. Moreover, the world was sensed by the system as a state of its range finder that changed in a continuous way. What the agent in fact had to accomplish was to form its own categorization of these sensory states. Analysing this model in the same way as NETtalk, the predefined components could be characterized as 'when the system bumps into an obstacle it turns away from it' and 'when the system detects a target it will turn towards it'. These reflexes were also precoded in terms of specific input–output relations, in this case between the US groups and the UR group. These precodings were not symbolic, but were expressed in a causal relationship between sensors, mediating structures and actuators. Such causal relations are in the domain of dynamics. The representations we can ascribe to the system do not relate to these predefined input–output mappings, but to the way in which the distal sensor is integrated into action. This integration process is driven by the dynamic relation between the phenotype, the predefined values, and the properties of the environment. By adapting to this system–environment interaction, however, the importance of the designer-dependent reflexive behaviours diminishes. Ultimately the reflexive actions will be completely replaced by learned actions. Moreover, the values around which the agent was constructed are all expressed in structural terms, which makes them directly testable in a biological domain.

The observed behavioural strategies, which could be called 'pulling out of an impasse' and 'wall following', relate to a time span of a large number of actions. The system itself, however, acts only on the basis of immediate sensory states: it is reactive. Only when we decompose the emergent behavioural sequences into their constituting local actions, the individual movements, can their relation to the properties of the agent and the environment be assessed. These properties do not contain the observed macroscopic behavioural patterns that the agent exhibits. These emergent properties have a *temporal* order that is not encoded in the control structure. Therefore, the observed emergent properties constitute a level of description that is not related to the properties implemented in the system. They could not be predicted from them. These sequences of actions, behaviours, are at the interface between the agent and its environment and they are ascribed to the agent by the observer.

The symbolic task description which was transferred to NETtalk limits its performance. NETtalk can never learn to pronounce words that are not described in the set of phonetic features used. The central aspect here is

that in these models the relationship between input and output is predefined at the most explicit level possible. One discrete and unique input state is associated with one unique output state: for example, 'pronounce input letter "a" with output features "central 2", "low", and "tensed"'. These systems can only master a task if it is symbolically mastered by its designer: if the input and output states and the 'knowledge' that connects input and output is properly predefined. These systems are in no way able to exceed the predefined domain of 'knowledge' through learning. Moreover, they are not robust: they will break down in any situation that was not foreseen by their designers (a well-known problem of AI). This shows a strong parallel to the way in which learning is defined in traditional AI (see also [Verschure and Pfeifer 1993]). Another parallel with traditional AI is that one assumes that these representational primitives can be reliably transduced from the sensors and to the actuators.

The predefined elements used in DAC (i.e. the value scheme) also limit the behavioural potential of the system. This limitation, however, has a completely different nature than the predefined elements used in the discussed subsymbolic models. In DAC the predefinitions limit the perceptual and behavioural potential of the system: for example, if an agent only has a range finder it can never deal with colour. The 'knowledge' that connects sensing to acting, however, is not predetermined. The system has to figure out itself which subset of the sense–act potential is useful to fill in, given its interaction with the environment: for example, how the range finder is integrated in the action system. This filling in is dependent on the phenotype, the value system, and the relation to the environment.

The DAC design principles give the system an *inherent* robustness. This is also demonstrated by the way in which the presented control structure solves problems in navigation tasks. In contrast, designers of traditional (symbolic) robot control architectures spend a lot of time defining all details of the input–output mapping that must be present in a system to enable it to deal with simple navigation tasks (see [Verschure *et al.* 1992] for the DAC solution to local minima in navigation tasks). This robustness is also demonstrated by the ease with which this control structure can be used in different experimental environments using simulations or robots (see [Mondada and Verschure 1993], its behaviour in parameter space [Almassy and Verschure 1992], its capacity for generalization [Verschure and Pfeifer 1993], and secondary conditioning [Verschure and Coolen 1991].

7.15 Solving the mind–brain dilemma

It is rather ambitious claiming to be able to solve the mind–brain dilemma that has been haunting cognitive science from its beginning. Subsymbolic connectionism presupposed that the gap between mind and brain could be closed by assuming that a subsymbolic level of description could mediate

between a symbolic and a nonsymbolic one. The evidence presented, however, indicates that this strategy did not bring the understanding of cognition much closer to the brain. They seemed to have lost their way in a symbolic trap.

I propose another approach to the problem of complete reduction. As has already been indicated by many others (e.g. [Maturana 1970], [Dennett 1978] and [Clancey 1989]), there are different stances to describing complex systems like the brain. While cognitivists prefer to stick to a symbolic description of behavioural regularities, neurobiologists adopt a level of description that has been called 'implementational'. Each of these groups of scientists have their own pragmatic reasons to believe that the level of description to which they adhere gives them the conceptual tools they need to study the phenomena in which they are interested. Subsymbolic connectionism seems to be founded on the belief that the regularities captured in symbolic descriptions of a task domain reflect regularities at the level of implementation. With the given examples from the experiments within the DAC paradigm I have tried to show that this assumption might not be generally valid. Observed behavioural regularities, which can be symbolically labelled, do not have to relate to identical regularities in the dynamics of the causal structure generating them. This implies that the reduction proposed by subsymbolic connectionism is one that might not be very feasible. It seems that the status of these different levels of description has to be redefined.

An alternative interpretation of the relation between a symbolic and a neurobiological approach, which might be more compatible with the reality of the phenomena we are studying in cognitive science, is that this relation is only present in the eye of the beholder and not in the system that is actually displaying the behaviour under investigation (e.g. [Maturana 1970], [Braitenberg 1984] and [Clancey 1989]). This means that the mind–brain dilemma and the problem of complete reduction are artifacts of the different conceptualizations that we, the observers of behaviour, have developed to understand cognition and not one that is actually present in the nature of the systems displaying this behaviour. The analysis presented of the DAC model illustrates this point. The behavioural regularities observed could only be explained in terms of the dynamics of the ongoing interaction of an agent and its environment. This coherence is lost when one decomposes these behavioural sequences into their basic actions. What, however, constitutes a coherent behavioural sequence is defined by the observer and not by the acting system itself. In the example discussed the system will just avoid or approach according to its immediate sensory states. The sensory states just before or after the executed action are of little relevance to it.

When we place NETtalk into this perspective this would mean that the symbolic precodings should be placed outside the system, back in the real world. They define the behavioural regularities that should be produced by

the model, not its internal dynamics. The task itself should be performed by a system consisting of realistic sensors and effectors with which it then has to learn to pronounce English text. Only when this system consisting of artificial eyes, ears, vocal tract and mouth is able to do the job will the claim of emergence in this model be substantiated. Of course, the pronunciation features that are predefined in NETtalk are one of the first things that the system should learn, next to the ability to categorize its sensors in such a way that characters and words can be perceived. This, however, would imply that the subsymbolic dream would be shattered anyway, since this system would not rely on symbols or subsymbols, but only on the transformation of states of symbolically unlabelled neurons in its visual and/or auditive systems to activation in its vocal tract and mouth. Only later on could the produced regularities be labelled by an observer in symbolic (or subsymbolic) terms and tested against the relevant behavioural data (like the observation of linguists that vowels are pronounced with a different set of speech actions than consonants). One could argue that since we know that with a limited set of phonetic features we can describe the regularities in speech production we might as well use it in our explanation of this phenomenon. We should take into account, however, that speech is a dynamic process. The categories we use to describe its regularities can vary over time (e.g. [Eimas *et al.* 1987]). If we rely on systems with a preset number of pronunciation features with static properties we will never be able to address the properties of speech as it occurs *in vivo* where genetic predefined processes are further adapted to the richness of the linguistic environment in which we find ourselves.

It is in the relation between observed behavioural regularities and the internal dynamics of control structures that emergence becomes a relevant issue. The concept of emergence is a difficult one. I will not try to come up with a complete definition here but will limit myself to the suggestion that any definition of emergence in the context of models that are developed to study cognition should include time as a basic dimension. Behavioural regularities that can be symbolically labelled are expressed in a temporal domain. For instance, the pronunciation features precoded in NETtalk are induced from sounds that can only become spoken words in a temporal domain. The DAC example illustrated that observed behavioural regularities, however, can consist of many local actions. Only through the system–environment interaction do these disconnected local actions become organized patterns of behaviour that look meaningful and coherent to an observer. These coherent behavioural patterns do not have to be mirrored in the dynamics of the control structure. The internal dynamics and the emergent behavioural regularities function at different time scales. Therefore, it becomes crucial to study behaviour not only in an embodied manner but also in a temporal perspective. Otherwise, the connection to the level

of behaviour, which can sometimes be symbolically labelled, will be impossible to make.

Another example of the different time domains in which observed ordered behaviour and local dynamics take place is provided by our work on classical conditioning [Verschure and Coolen 1991]. As mentioned earlier, this model took the Rescorla and Wagner laws of classical conditioning, which describe the development of associations between conditioned stimuli and conditioned responses over time, as its behavioural constraints. One of these constraints, blocking, is interpreted by Rescorla and Wagner as 'systems only learn when events violate their expectations'. This phenomenon can only be observed over a relatively long period of time. In our model we showed that this behaviour can be explained by local properties of a learning rule which functions in a much shorter time frame where potentiation and depression were tied to the competition between the plastic connections. It is not necessary to explain blocking in intentional terms, like 'expectation' or other abstract psychological constructs. Small-scale non-symbolic local dynamics suffices.

It seems that the only way to make some progress is not to try to force a decision between a symbolic approach and a non-symbolic one (as, for instance, critics of connectionism such as [Fodor and Pylyshyn 1988] propose), but to try to assess how each of these levels of description can contribute to our understanding of cognition and to develop models which bring both levels of description together. Such an integration does not require that one framework be reduced to the other, but acknowledges that symbolic accounts may provide valuable and necessary top-down constraints on the dynamical (connectionist) systems that can be constructed. The brain provides the bottom-up constraints necessary to validate the proposed control structure, and in this way provides insight into the primitives of the internal dynamics that generate the behavioural regularities we observe. Moreover, by concentrating on properties of the brain (e.g. on local dynamics to perform the work and on the importance of distributed processing) we can be protected from postulating too much ungrounded knowledge in our models. The control structures, expressed in connectionist terms, which incorporate these constraints will allow us to explore the principles of the most sophisticated system we know that generates behaviour in its interaction with the real world: the brain. This strategy establishes an ongoing interaction between the levels of description involved. In [Verschure and Pfeifer 1993] there is presented a concrete analysis that follows this strategy, relating Edelman's extended theory of neuronal group selection to Newell's exemplar unified theory of cognition, SOAR.

Solving the mind–brain dilemma requires an approach devoid of dogmas and bandwagons in order to avoid a further fragmentation of the field. What is needed is a global perspective which encompasses the multitude of disciplines and descriptive levels of relevance, from genetics and molecular

biology, with its emphasis on morphogenesis, to anthropology and linguistics, relating to the cultural and linguistic environment in which we are thrown. Although the main ambition of cognitive science has been to support and develop such an interaction it has not been realized. One of the reasons for this situation has been that the models and theories that have been developed were enslaved by the ontology of their designers, like the examples analysed which stem from the subsymbolic school of connectionism. A move from the Platonic world of rules and representations to the real world of the dynamics of system–environment interaction seems appropriate.

7.16 Discussion

Subsymbolic connectionism has presented itself as a new theory of the mind that will enlighten our view of age-old questions like the mind–body problem. By analysing paradigmatic examples of this approach it was demonstrated that the explanational scheme it proposes is too problematic to be acceptable. In the case of subsymbolic connectionism the issue of levels seems to lead to a confusion of levels. At the heart of this problem lies the mind–brain dilemma: the traditional conflict between symbolic and non-symbolic approaches towards understanding the mind and the brain.

It was shown that the proper relation between symbolic characterizations of behaviour and the internal dynamics of the control structures that generate it can only be established when we place both in a temporal domain. In doing so, symbolic descriptions of observed macroscopic behavioural regularities can be related to sequences of internal microscopic dynamical processes. We do not have to assume that these internal processes somehow mimic the systematicity of the symbolic constructs that we as observers of behaviour have ascribed to the behaving system.

Subsymbolic connectionists seem to believe that this mimicking does take place. By analysing NETtalk I tried to show that this implied that they were in fact applying a connectionist technique in the traditional symbolic paradigm. In no way did this move settle the mind–brain dilemma. An alternative approach would be to place the observed behavioural regularities outside the system. In the case of NETtalk this means that articulatory features are extracted by an observer from overt behaviour, not from internal mechanisms. These symbolic descriptions of macroscopic behavioural regularities provide valuable guidelines in the search for the mechanisms generating this behaviour. Any model that is supposed to explain psychological processes must be validated against the behavioural regularities captured by symbolic descriptions. We have to acknowledge, however, that the observed behaviour has to be placed at the interface between a behaving organism and its environment. In any analysis these two components have to be included. In trying to understand behaviour we cannot assume

that there exists a mirroring relationship between the observed behavioural regularities and the structure and dynamics of the mechanisms that underlie it. Therefore, in the attempt to explain control structures like the brain the (neuro)biological constraints are as important as the behavioural ones. Only by bringing these two sources of constraints together can we find complete characterizations of the internal mechanisms that generate behaviour.

Two standard objections to such a synthetic approach towards cognition are that it is unclear how it can ever account for complicated psychological processes like problem solving, concept formation and language, and that it resembles behaviourism.

The first objection does not constitute a principled problem. We must be very clear about the status of present theories in this domain. As has already been indicated, cognitivism has not produced any solid answers here and is still facing a number of very serious problems. The only solution to these problems can be found in relating symbolic descriptions to behaving systems and the control structures that drive them. In this way a connection to the necessary additional constraints from the domain of (neuro)biology can be established. The basic problem here is that the symbolic conceptualizations of these phenomena and the way neural mechanisms that seem to relate to them are described do not seem very compatible. If we agree that connecting symbolic descriptions of behaviour to neural dynamics makes sense, then this also implies that we must be prepared to reconceptualize our macroscopic descriptions. For the study of language, this implies that it has to be placed in a biological setting and not just in the abstract domain of computations. This step towards biology can provide new insights. As an example we might consider the work by [Edelman 1989] who demonstrates how, from a biological point of view, there only needs to be one basic epigenetic mechanism for the acquisition of language which is semantic in nature. This mechanism is founded in the capacity of the brain to categorize in an expanding time window. This suggestion relates to the hypothesis of semantic bootstrapping proposed by [Pinker 1984]. The basic distinction between syntax and semantics, which has been the foundation of modern linguistics, can be dropped. Syntax is acquired by interacting in a linguistic environment: it arises out of semantics.

In [Verschure, 1994] a further generalization of the principles expressed in DAC towards the domain of representing sequences of sense–act relationships is presented. It is shown how a simulated agent can indeed bootstrap itself to a qualitatively new level of representation by exploiting the macroscopic regularities of its interaction with the world. This system not only adaptively builds up its 'body of knowledge' but also develops sequences, and makes recombinations of their components. This agent also shows an improvement in its behaviour as compared with a purely reactive one. Although this model is a first exploration of this domain it does show that purely bottom-up principles can give rise to higher-order representations.

This implies that the classical argument against associationist approaches towards these issues – that it is not feasible how they can give rise to higher-order cognition – has lost its validity. It is not only feasible, it has been demonstrated.

The second objection is not valid. One of the basic characteristics of the type of behaviourism which is referred to in this case is that it excluded internal mechanisms from its agenda. This was a reaction to the psychological tradition of phenomenology which completely relied on subjectivistic interpretations of behaviour. In the approach proposed here the internal mechanisms that mediate between sensing and acting are considered crucial in understanding cognition. The example from the DAC paradigm illustrates this point. This objection is better directed against proponents of the cognitivistic tradition (e.g. [Fodor 1984]) who assumed that we need not investigate actual mechanisms and 'central' processes. In contrast to behaviourism, however, the reason here is not to keep our conceptualizations as objective and controllable as possible, but to place the processes that have to account for the computational structure of cognition outside the field of empirical validation. This effectively means that in the study of behaviour we can, according to the thesis of modularity, only deal with peripheral modules like perception and should forget about internal processes.

Connectionism has opened up a new and very promising field of research in cognitive science by re-emphasizing the importance of the brain in studying the mind. If, however, we allow the promise of combining psychology and biology to deteriorate into the application of connectionist methods in the traditional cognitivistic paradigm we should not be too surprised if the whole endeavour slowly dies away. Connectionism can only become part of a new paradigm for the study of the mind, brain and behaviour when it appreciates the virtue of dynamics above computation. This means that instead of breaking cognition down into the knowledge a system must possess we should try to trace it back to its fundamental adaptational mechanisms. This move towards dynamics will place symbolic characterizations outside the system, and behaviour back in the temporal domain of system–environment interaction.

Acknowledgements

This chapter was written while the author was working at the AI lab of the University of Zurich. He is very much indebted to Dom Massaro, Dean Allemang and Markus Stolze for their helpful comments and fruitful discussions. An early version of the analysis put forward in this chapter was presented in 1990 at a conference as part of the 'Mind and Brain' project at the Zentrum für interdisziplinäre Forschung (ZiF) in Bielefeld, Germany.

7.17 Appendix A: The description of the model in section 7.7

In this one-layered model (or autoassociator) every element i can take on a continuous activation value a_i which ranges between -1 and 1. All elements are connected to each other. Every connection between unit i and unit j has a weight, w_{ij}, which modulates the transmitted signal. Every unit i receives input from two sources. The first, external, source is determined by the pattern the model has to learn. Every unit i has clamped on it an external input value, e_i, which is determined by element i of the input pattern. The other source of input is internal and is determined by the transmitted activations of the other units and the connecting weights in the model. For every unit i this internal input, int_i, is determined by the sum of the weighted signals of the other j units (in contrast to the McClelland and Rumelhart implementation in which self-connections are allowed):

$$\text{int}_i = \sum_{j=1}^{N} a_j w_{ij} \tag{7.1}$$

The change of activation of unit i is determined by the total input, net_i, to unit i.

$$\text{net}_i = \text{int}_i + \text{ext}_i \tag{7.2}$$

The activation a_i of element i at time step $t+1$ is determined by the activation at moment t and net_i:

$$a_i(t+1) = a_i(t) + \Delta a_i(t) \tag{7.3}$$

$$\Delta a_i(t) = E\text{net}_i(1 - a_i(t)) - Da_i(t) \quad \text{if net}_i > 0 \tag{7.4}$$

$$\Delta a_i(t) = E\text{net}_i(-1 - a_i(t)) - Da_i(t) \quad \text{if net}_i \leq 0 \tag{7.5}$$

E and D, respectively, define excitation and decay. The strengths of the associations between the units develop according to

$$w_{ij}(t+1) = w_{ij}(t) + \Delta w_{ij}(t) - Cw_{ij} \tag{7.6}$$

where C denotes the decay of the weights. The change of the weights $\Delta w_i j$ is dependent on the error s_i (between the actual activation of unit i and the expected activation e_i) and the activation of unit j.

$$\Delta w_{ij}(t) = \eta \sigma_i a_j(t) \tag{7.7}$$

where η denotes the learning rate parameter and σ_i is given by:

$$\sigma_i = \text{ext}_i - \text{int}_i \tag{7.8}$$

7.18 References

Ackley, D.H., Hinton, G.E., Sejnowski, T.J.: A learning algorithm for Boltzmann machines, *Cognitive Science*, **9**, 147–169, 1985.

Agre, P.E., Chapman, D.: Pengi: an implementation of a theory of activity, *Sixth National Conference on Artificial Intelligence*, Seattle, WA, pp. 268–272, 1987.

Almassy, N., Verschure, P.F.M.J.: Optimizing self-organizing control architectures with genetic algorithms: the interaction between natural selection and ontogenesis. in: Männer R., Manderick B., *Proceedings of the Second Conference on Parallel Problem Solving from Nature*, Amsterdam: Elsevier, pp. 451–460, 1992.

Amit D.J.: *Modelling Brain Function: The World of Attractor Neural Networks*, New York: Cambridge University Press, 1989.

Anderson J.A., Murphy G.L.: Concepts in connectionist models. in Denker J.S. (ed.): *Neural Networks for Computing: AIP Conference Proceedings*, New York: American Institute of Physics, pp. 17–22, 1986.

Bechtel W.: Connectionism and intentionality, *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society*, Hillsdale NJ: Erlbaum, pp. 553–600, 1989.

Beer R.D.: *Intelligence as Adaptive Behavior: An Experiment in Computational Neuroethology*, New York: Academic Press, 1990.

Braitenberg V.: *Vehicles: Experiments in Synthetic Psychology*, Cambridge, MA: MIT Press, 1984.

Broadbent D.: A question of levels: comments on McClelland and Rumelhart, *Journal of Experimental Psychology* **114**, 189–192, 1985.

Brooks R.A.: Intelligence without reason, *Proceedings of the Twelfth International Conference on Artificial Intelligence (IJCAI)*, vol 1, San Mateo, CA: Morgan Kaufmann, pp. 569–595, 1991a.

Brooks R.A.: Intelligence without representation, *Artificial Intelligence* **47**, 139–159, 1991b.

Changeux J.P.: *Neuronal Man: The Biology of Mind*, Oxford University Press, 1985.

Chapman D.: Planning for conjunctive goals, *Artificial Intelligence* **32**, 333–337, 1987.

Chomsky N.: A review of B.F Skinner's *Verbal Behavior*, *Language* **35**, 26–58, 1959.

Churchland P.M.: *The Neurocomputational Perspective*, Cambridge, MA: MIT Press, 1989.

Churchland P.S, Sejnowski T.J.: Neural representation and neural computation, in: Galaburda A.M. (ed.): *From Reading to Neurons*, Cambridge, MA: MIT Press, 1989.

Clancey W.J.:. The frame of reference problem in cognitive modeling, *Proceedings of the Annual Conference of the Cognitive Science Society*, Hillsdale, N.J.: Lawrence Erlbaum, pp. 107–114, 1989.

Clancey W.J.: The frame of reference problem in the design of intelligent machines, in Lehn K. van (ed.): *Architectures for Intelligence, Proc. 22nd Carnegie Symposium on Cognition*, Hillsdale, NJ: Erlbaum, pp. 357–423, 1992.

Clark A.: *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing*, Cambridge, MA: MIT Press, 1989.

Crick F.: The recent excitement about neural networks, *Nature* **337**, 129–132, 1989.

Dennett D.C.: *Brainstorms: Philosophical Essays on Mind and Psychology*, Cambridge, MA: MIT Press, 1978.

Edelman G.M.: *Neural Darwinism: The Theory of Neuronal Group Selection*, New York: Basic Books, 1987.

Edelman G.M.: *The Remembered Present: A Biological Theory of Consciousness*, New York: Basic Books, 1989.

Edelman G.M.: *Bright Air, Brilliant Fire: On the Matter of the Mind*, New York: Basic Books, 1992.

Eimas P.D., Miller J.L., Jusczyk P.W.: On infant speech perception and the acquisition of language, in Harnad S. (ed.): *Categorical Perception: The Groundwork of Cognition*, Cambridge University Press, pp. 161–195, 1987.

Estes W.K.: Toward a framework for combining connectionist and symbol-processing models, *Journal of Memory and Language* **27**, 196–212, 1988.

Farde L., Wiesel F., Halldin C., Stone-Elander S., Sedvall G.: No D2 receptor increase in a PET study of schizophrenia, *Archives of General Psychiatry* **44**, 671–672, 1987.

Farmer J.D.: A Rosetta Stone for connectionism, *Physica D* **42**, 153–187, 1990.

Fodor J.A.: Special sciences, or the disunity of science as a working hypothesis, *Synthese* **28**, 97–115, 1974.

Fodor J.A.: *The Modularity of the Mind*, Cambridge, MA: MIT Press, 1984.

Fodor J.A., Pylyshyn Z.W.: Connectionism and cognitive architecture, a critical analysis, *Cognition* **28**, 3–71, 1988.

Furth H.G.: *Piaget and Knowledge: Theoretical Foundations*, University of Chicago Press, 1969.

Gardner H.: *The Mind's New Science: A History of the Cognitive Revolution*, New York: Basic Books, 1987.

Gorman R.P., Sejnowski T.J.: Analysis of hidden units in a layered network trained to classify sonar targets, *Neural Networks* **1**, 75–89, 1988.

Harnad S.: The symbol grounding problem, *Physica D* **42(1–3)**, 335–346, 1990.

Haugeland J.: The nature and plausibility of cognitivism, *Behavioral and Brain Sciences* **2**, 215–260, 1978.

Hebb D.O.: *The Organization of Behavior*, New York: Wiley, 1949.

Hinton G.E., Plaut D.C., Shallice T.: Lesioning an attractor network: investigations of acquired dyslexia, *Psychological Review* **98**, 74–95, 1991.

Hinton G.E., Plaut D.C., Shallice T.: Simulating brain damage, *Scientific American*, October, 58–65, 1993.

Hodgkin A.L., Huxley A.F.: A quantitive description of membrane current and its application to conduction and excitation in nerve, *Journal of Physiology* **117**, 500–544, 1952.

Hofstadter D.R.: *Metamagical Themas*, Harmondsworth, Middlesex: Penguin, 1985.

Hornik K., Stinchcombe M., White H.: Multilayer feedforward networks are universal approximators, *Neural Networks* **2**, 359–366, 1989.

Hubel D.N., Wiesel T.N.: Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, *Journal of Physiology* **160**, 106–154, 1962.

Hubel D.N., Wiesel T.N.: Receptive fields and functional architecture of monkey striate cortex, *Journal of Physiology* **195**, 215–243, 1968.

Hull C.L., Baernstein H.D.: A mechanical parallel to the conditioned reflex, *Science* **70**, 14–15, 1929.

Klatt D.H.: Review of text-to-speech conversion for English, *Journal of the Acoustical Society of America* **82**, 737–793, 1987.

Kosslyn S.M., Chabris C.F., Koenig O.: Categorical versus spatial relations: computational analyses and computer simulations, *Journal of Experimental Psychology: Human Perception and Performance* **18**, 562–577, 1992.

Kupfermann I., Weiss K.R.: The command neuron concept, *Behavioral and Brain Sciences* **1**, 3–39, 1978.

Lee D.N.: A theory of visual control of breaking based on information about time to contact, *Perception* **5**, 437–459, 1976.

Malcolm C., Smithers T., Hallam J.: An emerging paradigm in robot architecture, in Kanade T., Groen F.C.A., Herzberger L.O. (eds.): *Intelligent Autonomous Systems 2*, Amsterdam: Elsevier, 1989.

Massaro D.W.: Some criticisms of connectionist models of human performance, *Journal of Memory and Language* **27**, 213–234, 1988.

Massaro D.W.: Testing between the TRACE model and the fuzzy logical model of speech perception, *Cognitive Psychology* **21**, 398–421, 1989.

Maturana H.R.: Biology of cognition, reprinted in Maturana H.R., Varela F.: *Autopoiesis and Cognition: The Realization of the Living*, Dordrecht: Reidel, 1980.

McCarthy J., Hayes P.J.: Some philosophical problems from the standpoint of artificial intelligence, in Meltzer B., Michie D. (eds.): *Machine Intelligence* **4**, pp. 463–502, 1969.

McClelland J.L., Rumelhart D.E. *et al.*: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 2: Psychological and Biological Models*, Cambridge, MA: MIT Press, 1986.

McCloskey M.: Networks and theories: the place of connectionism in cognitive science, *Psychological Science* **2**, 387–395, 1991.

Miklos G.L.G.: Molecules and cognition: the latterday lessons of levels, language, and Iac. Evolutionary overview of brain structure and function in some vertebrates and invertebrates, *Journal of Neurobiology* **24**, 842–890, 1993.

Miller N.E.: Liberalization of basic S-R concepts: extensions to conflict behavior, motivation, and social learning, in Koch S. (ed.): *Psychology: A Study of a Science*, Vol. 2, New York: McGraw-Hill, pp. 196–202, 1959.

Minsky M., Papert S.: *Perceptrons, An Introduction to Computational Geometry*, Cambridge, MA: MIT Press, 1969.

Mitchell M., Hofstadter D.R.: The emergence of understanding in a computer model of concepts and analogy-making, *Physica D* **42**, 322–334, 1990.

Mondada F., Verschure P.F.M.J.: Modeling system–environment interaction: the complementary roles of simulations and real world artifacts, *Proceedings of the Second European Conference on Artificial Life*, pp. 808–817, 1993.

Moore E.F.: Gedanken-experiments on sequential machines, in Shannon C.E., McCarthy J. (eds.): *Automata Studies*, Princeton University Press, pp. 129–153, 1956.

Newell A.: The knowledge level, *AI magazine* **2**, 1–20, 1981.

Newell A.: *Unified Theories of Cognition*, Cambridge, MA: Harvard University Press, 1990.

Newell A., Simon H.A.: GPS, a program that simulates human thought, in Feigenbaum E.A., Feldman J. (eds.): *Computers and Thought*, New York: McGraw-Hill, pp. 279–296, 1963.

Newell A., Simon H.A.: Computer science as empirical inquiry: symbols and search, *Communications of the ACM* **19(3)**, 113–126, 1976.

Piaget J.: *Biology and Knowledge*, University of Chicago Press, 1971.

Piatelli-Palmarini M. (ed.): *Language and Learning, The Debate between Jean Piaget and Noam Chomsky*, Cambridge, MA: Harvard University Press, 1980.

Pinker S.: *Language Learnability and Language Development*, Cambridge, MA: Harvard University Press, 1984.

Pinker S., Prince A.: On language and connectionism: analysis of a parallel distributed processing model of language acquisition, *Cognition* **28**, 73–193, 1988.

Pomerleau D.A.: ALVINN: an autonomous land vehicle in a neural network, in Tourezky D.S. (ed.): *Advances in Neural Information Processing Systems 1*, San Mateo CA: Morgan Kaufmann, pp. 305–313, 1989.

Putnam H.: Minds and machines, in Hook S. (ed.): *Dimensions of Mind*, New York University Press, 1960.

Pylyshyn Z.: *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*, Norwood, NJ: Ablex, 1987.

Pylyshyn Z.: Computing in cognitive science, in Posner M.I. (ed.): *Foundations of Cognitive Science*, Cambridge, MA: MIT Press, pp. 51–91, 1989.

Ramsey W., Stich S.P., Garon J.: Connectionism, eliminativism, and the future of folk psychology, in Ramsey W., Stich S.P., Rumelhart D.E. (eds.): *Philosophy and Connectionist Theory*, Hillsdale, NJ: Lawrence Erlbaum, pp. 199–228, 1991.

Rescorla R.A., Wagner A.R.: A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement, in Black A.H., Prokasy W.F. (eds.): *Classical Conditioning 2, Current Theory and Research*, New York: ACC, pp. 64–99, 1972.

Rosenberg C.: Revealing the structure of NETtalk's internal representations, *Proceedings of the Ninth Annual Conference of the Cognitive Science Foundation*, Hillsdale NJ: Lawrence Erlbaum, pp. 537–554, 1986.

Rosenberg C., Sejnowski T.: Parallel networks that learn to pronounce English text, *Complex Systems* **1**, 145–168, 1987.

Rosenblatt F.: The Perceptron: a probabilistic model for information storage in the brain, *Psychological Review* **65**, 386–408, 1958.

Rosenblatt F.: *Principles of Neurodynamics*, Washington: Spartan, 1962.

Rumelhart D.E., Hinton G.E., Williams R.J.: Learning representations by back propagating errors, *Nature* **323**, 533–536, 1986.

Searle J.R.: Minds, brains, and programs, *Behavioral and Brain Sciences* **3**, 417–451, 1980.

Seidenberg M.S., McClelland J.L.: A distributed, developmental model of word recognition and naming, *Psychological Review* **96**, 523–568, 1989.

Sejnowski T.J., Koch C., Churchland P.S.: Computational neuroscience, *Science* **241**, 1299–1306, 1988.

Sejnowski T.J., Rosenberg C.R.: NETtalk: a parallel network that learns to read aloud, The Johns Hopkins University Electrical Engineering and Computer Science technical report 86/01, 1986.

Simon H.A.: *Sciences of the Artificial*, Cambridge, MA: MIT Press, 1969.

Smith B.C.: The owl and the electric encyclopedia, *Artificial Intelligence* **47**, 251–288, 1990.

Smolensky P.: The constituent structure of connectionist mental states: a reply to Fodor and Pylyshyn, *The Southern Journal of Philosophy* **25**, sup., 1987.

Smolensky P.: On the proper treatment of connectionism, *Behavioral and Brain Sciences* **11**, 1–73, 1988.

Suchman L.A.: *Plans and Situated Actions, The Problem of Human/Machine Communication*, Cambridge University Press, 1987.

Sutton R.S., Barto A.G.: Toward a modern theory of adaptive networks: expectations and prediction, *Psychological Review* **88(2)**, 135–170, 1981.

Uttal W.R.: On the two-way barriers between models and mechanisms, *Perception and Psychophysics* **48**, 188–203, 1990.

Verschure P.F.M.J.: Taking connectionism seriously: the vague promise of subsymbolism and an alternative, *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*, Hillsdale, NJ: Erlbaum, pp. 653–658, 1992.

Verschure P.F.M.J.: Formal minds and biological brains: AI and Edelman's extended theory of neuronal group selection, *IEEE Expert: Intelligent Systems and Their Applications.* **8(5)**, 66–75, 1993.

Verschure P.F.M.J.: The cognitive development of an autonomous artifact: the self-organization of categorization, sequencing, and chunking, Working paper ZiF workshop on Prerational Intelligence, Zentrum für interdisziplinäre Forschung, Bielefeld, 1994.

Verschure P.F.M.J., Coolen A.C.C.: Adaptive fields: distributed representations of classically conditioned associations, *Network* **2**, 189–206, 1991.

Verschure P.F.M.J., Pfeifer R.: Categorization, representations, and the dynamics of system–environment interaction: a case study in autonomous systems, in Meyer J.A., Roitblat H., Wilson S. (eds.): *From Animals to Animats: Proceedings of the Second International Conference on Simulation of Adaptive Behavior*, pp. 210–217, 1993.

Verschure P.F.M.J., Kröse B.J.A., Pfeifer R.: Distributed adaptive control: the self-organization of structured behavior, *Robotics and Autonomous Systems* **9**, 181–196, 1992.

Willshaw D.J., Von der Malsburg C.: How patterned neural connections can be set up by self-organization, *Proceedings of the Royal Society of London* **B 194**, 431–445, 1976.

Winograd T., Flores F.: *Understanding Computers and Cognition: A New Foundation for Design*, Norwood, NJ: Ablex, 1986.

Wong D.F. *et al.*: Positron emission tomography reveals elevated D2 dopamine

receptors in drug-naive schizophrenic patients, *Science*, 1558–1563, 1986.

Wray J., Green G.G.R.: How neural networks work: the mathematics of networks to solve standard engineering problems, *Proceedings of American Control Conference*, 2311–2314, 1991.